Kenny, D. A., Kashy, D., & Bolger, N. (1998). Data analysis in social psychology. In D. Gilbert, S. Fiske, and G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., pp. 233-265). New York: McGraw-Hill.

DATA ANALYSIS IN SOCIAL PSYCHOLOGY

DAVID A. KENNY, *University of Connecticut*
DEBORAH A. KASHY, *Texas A & M University*
NIALL BOLGER, *New York University*

Since the state of data analysis in social psychology was last reviewed in the *Handbook of Social Psychology* (Kenny, 1985), there have been many important advances in data-analytic methods designed to address problems specific to the study of human social behavior. These new techniques are replacing the data-analytic approaches that were initially developed for agricultural research. Problems such as nonindependence of observations, measurement error, and generalizability of results from specific operations are being directly addressed.

Among these recent developments, meta-analysis represents one of the most important methodological advances in the social and behavioral sciences over the last 25 years. Its capacity to integrate rigorously the results of multiple studies has already proven invaluable in a myriad of substantive areas. Because several sources detailing the methods of meta-analysis already exist, we do not discuss this method in this chapter. We refer the interested reader to an excellent review by Cooper and Hedges (1994).[1]

Similarly, major strides have been made in the analysis of data in which persons interact with or rate multiple partners. For the analysis of nominal outcomes (e.g., sociometric judgments of liking), Wasserman and Faust (1994) provide an almost encyclopedic coverage. For the analysis of variables measured at the interval level of measurement, Kenny (1996a) details recent designs, models, and analysis techniques.

Despite advances such as these, the standard data-analytic tool for most social psychologists remains the analysis of variance (ANOVA). We use ANOVA so much in our thinking that we have wondered whether laypeople also use ANOVA to make sense of their world (Kelley, 1967). Recent developments in ANOVA are presented in review articles by Wilcox (1987) and Judd, McClelland, and Culhane (1995) and in texts by Judd and McClelland (1989), Maxwell and Delaney (1990), and Harris (1994). Moreover, Abelson (1995) has written a thoughtful book on statistics, much of which covers ANOVA issues. The first part of the chapter is an extended discussion of the analysis of data from group research, largely from an ANOVA framework. Nonindependence of observations is a serious issue that is often just ignored. We consider the consequences of using person or group as the unit of analysis on Type I and II errors.

With some reluctance, social psychologists have begun to recognize the limitations of ANOVA and are turning to more general methods of data analysis that overcome these limitations. In our chapter, we focus extensively on two such methods, structural equation modeling and multilevel modeling.

Structural equation modeling has been increasingly applied in social psychological research, most notably attitude structure. Although there is not a thorough and readable discussion of this method, Loehlin (1992) provides a general introduction and Hayduk (1987) and Byrne (1994) provide useful introductions to the computer programs of LISREL and EQS, respectively. Finally, Bollen's book (1989) serves as a beneficial technical resource. Although structural equation modeling has delivered fewer theoretical insights than were initially promised (though there are

This research was supported in part by grants from the National Science Foundation (DBS-9307949) and the National Institute of Mental Health (RO1-MH51964). Brian Lashley and Cynthia Mohr were helpful in preparing the chapter, and Charles Judd, Thomas Malloy, and Irene Elkin provided useful comments that helped us in revising the chapter. Robert Calsyn generously shared with us his data.

notable exceptions, e.g., Jussim, 1991), it has provided important clarifications in the conceptual meaning of measures (see Judd & McClelland, 1998, in this *Handbook*).

Despite the increasing use of this technique, there is still considerable confusion regarding two fundamental issues: (a) whether a given structural equation model can be estimated, a topic called identification; and (b) whether the results of a structural equation model indicate the existence of causal mediation. Determining whether a model is identified is, we feel, the least understood aspect of structural equation modeling among social psychologists. For this reason we treat it in detail.

Determining whether a structural equation model shows causal mediation is better understood by social psychologists, but there are nonetheless many instances in the literature where tests of mediation are carried out incorrectly. Therefore, we provide a detailed discussion of the estimation and testing of mediational models by expanding and clarifying the analysis proposed by Baron and Kenny (1986) and Judd and Kenny (1981).

In addition to structural equation modeling, multilevel modeling has also emerged as a competitor to ANOVA in the last decade. For social psychologists who work with repeated-measures data, this method is useful because it overcomes many limitations of repeated-measures ANOVA. It can handle situations where the between-subject and within-subject independent variables are continuous variables and where there are missing data on the repeated measure. For social psychologists who do small group research, it is useful because it can easily handle the nonindependence of persons within groups and does not require equal numbers of persons in each group. More generally, multilevel modeling can be used to analyze any data that involve at least two levels of analysis (e.g., persons with repeated measurements within each person and groups with multiple persons within each group). Because this method is likely to be unfamiliar to most social psychologists, we cover it extensively.

We begin the chapter by discussing nonindependence of observations in group research. After considering ANOVA solutions, we discuss how multilevel models can be used to estimate many forms of grouped data. Finally, we discuss identification in structural equation models and the problem of testing mediation.

UNIT OF ANALYSIS IN GROUP RESEARCH[2]

Researchers studying small groups, relationships, or organizations face the difficulty of choosing the unit of analysis. This problem arises because of the hierarchical structure of the data: individuals are nested within groups such that each person is a member of one and only one group. (Kenny [1996a] considers designs in which persons are members of more than one group.) Typically the choice of the unit of analysis is between person and group. If person is used as the unit of analysis, the assumption of independent units is likely to be violated because persons within groups may influence one another (Kenny & Judd, 1986). Alternatively, if group (i.e., couple, team, or organization) is used as the unit of analysis, the power of the statistical tests is likely to be reduced because there are fewer degrees of freedom than there are in the analysis that uses person as the unit of analysis. In this section, we concentrate on categorical, not continuous, independent variables. In addition, we do not consider independent variables that are random and operate at the group level (Griffin & Gonzalez, 1995; Kenny & La Voie, 1985).

In discussing the ramifications of the two choices for the unit of analysis, an important distinction must be made among three types of independent variables (denoted as A): nested, crossed, and mixed. A nested independent variable occurs when groups are assigned to levels of the independent variable such that every member of a given group has the same score on A with some groups at one level of A and other groups at other levels of A. A crossed independent variable occurs when A varies within the group, with some group members in one level of A and other group members in the other level of A, but for all groups the group average for A is the same. A mixed independent variable shares characteristics of both nested and crossed independent variables in that it varies both between and within groups. When the independent variable is mixed, persons within the group may differ on A and group averages on A may differ from group to group.

Consider the case in which A is gender. Gender would be nested if all groups contained same-gender members; gender would be crossed if each group consisted of both women and men with the restriction that each group has the same gender ratio; and gender would be mixed if the gender ratio varied from group to group as when some groups are same-gender and some are mixed-gender. A mixed variable, which is likely to be a new concept to most social psychologists, can provide significant conceptual leverage but also presents data-analytic difficulties. The question of whether group or person should be the unit of analysis is considered separately for the three types of independent variables. Within this section of the chapter, it is assumed that there are an equal number of persons per group and that there are two levels of A. In the multilevel modeling section of the chapter, the assumption of equal group sizes is relaxed.

Nested Independent Variable

Imagine the following hypothetical study: a researcher investigates the effect of two types of problem-solving strategies on group-member motivation. The researcher forms twenty five-person groups, ten of which use strategy 1 and ten of which use strategy 2. The key features of this example are the 100 persons and twenty groups, the two treatment conditions, and each group being in only one level of the treatment.

These data can be analyzed within an ANOVA framework as presented in Table 1. There are three sources of variation in the nested design. There is variation due to the independent variable, strategy type, which is denoted as factor *A* in the table. One type of strategy may be more motivating on average than the other type of strategy. Second, within each level of *A* some groups may be more motivated than other groups (*G/A*). Finally, within the groups some persons may be more motivated than others (*S/G/A*).

Generally, the central question addressed in a study of this type is whether there is an effect of strategy type or *A*. There are three different choices of error term with which one can test the effect of factor *A*. These are group within *A* (*G/A*), person within group within *A* (*S/G/A*), and person within *A* ignoring group (*S/A*). The *S/A* error term (the pooling of the *S/G/A* and *G/A* sources) involves treating person as the unit of analysis and ignoring group. The pooled sum of squares is 108 + 160 or 268 and the pooled degrees of freedom are 18 + 80 or 98. So person within *A* (ignoring group) has a mean square of 2.735, and the resulting *F* test of the strategy main effect is 10.97 with 1 and 98 degrees of freedom. Note that this approach in which person is the unit of analysis (and group is ignored) is equivalent to treating the design as if it were a single-factor ANOVA design in which there are only two sources of variation, *A* and *S/A*.

If group is used as the unit of analysis, *G/A* is used as the error term, and the *F* test equals 5.00 with 18 degrees of freedom. If person is used as the unit but group effects are controlled, the error term is *S/G/A*, and the *F* test is 15.00 with 80 degrees of freedom. Thus these three different choices concerning the unit of analysis yield three different *F*s with three different degrees of freedom and three different error terms. The appropriate choice among these three analyses is dictated by the degree to which the data within the groups are related or nonindependent.

TABLE 1

ANOVA Source Table for the Nested Design Example with Twenty Groups of Five Persons

Source	*SS*	*df*	*MS*
Between Groups			
Strategy (*A*)	30	1	30
Group (*G/A*)	108	18	6
Within Groups			
Person (*S/G/A*)	160	80	2

$$\rho = \frac{MS_{G/A} - MS_{S/G/A}}{MS_{G/A} + [n-1]MS_{S/G/A}}$$

(where *n* is the number of persons per group)

$$\rho = \frac{6-2}{6+[4]2} = .29$$

$$F(18{,}80) = \frac{MS_{G/A}}{MS_{S/G/A}} = 3.00, p < .001, \text{two tailed}$$

Measuring Nonindependence: The Intraclass Correlation When person is used as the unit of analysis for the data in Table 1, the *F* test is two to three times as large as it is when group is the unit of analysis. Although it is desirable to have a healthy *F* ratio, an assessment of group effects is needed before group can be ignored and *S/A* can be used as the error term. Group effects occur if the scores of individuals within a group are more similar to one another than are the scores of individuals who are in different groups. Because it seems reasonable to believe that individuals in some groups may be more motivated than those in other groups, the measurement and statistical evaluation of group effects are required.

The standard measure employed to assess group effects is the *intraclass correlation* which is denoted as ρ. The intraclass correlation measures the correlation between two persons' outcomes who are both in the same group. So an intraclass correlation of .25 means that the correlation between the motivations of two persons who are in the same group is .25. Alternatively, the intraclass correlation can be viewed as the amount of variance in the persons' scores that is due to the group, controlling for the effect of *A*.

The standard measure of the intraclass correlation uses the mean squares from the ANOVA. The ingredients to the formula are the mean square for groups within *A* (*G/A*), the mean square for persons within groups within *A* (*S/G/A*), and the number of persons per group (*n*; see Table 1). The intraclass correlation can alternatively be estimated using correlational methods instead of ANOVA (Griffin & Gonzalez, 1995).

The intraclass correlation is like a product-moment correlation in that its upper limit is one. However, its lower limit is not always minus one. In general, its lower limit is $-1/(n-1)$ where *n* is the number of persons per group. So if *n* is twenty, then the intraclass can be no smaller than −.0526. Note that if *n* is two as in dyadic analysis, the lower limit is minus one. An example of a negative intraclass correlation may occur with married couples if one member experiences positive outcomes from the treatment (*A*) but his or her spouse experiences negative outcomes. Although negative intraclass correlations are relatively rare, they generally should be given serious consideration.

After the intraclass correlation has been estimated, it is tested for statistical significance by an *F* test. To create the *F* ratio, one places the larger of the two mean squares ($MS_{G/A}$ and $MS_{S/G/A}$) on the numerator and the smaller mean square is placed on the denominator. The degrees of freedom for *F* are determined accordingly. The obtained *F* is compared to a critical value for which the *p* value is divided by two. The *p* value is divided by two because, unlike the typical *F* test in ANOVA, both tails of the *F* distribution are being used, as in the use of the F distribution to test for unequal variances. A significant F statistic implies that there is nonindependence of data within the groups.

Because the intraclass correlation is used in determining whether there is nonindependence in the data, it is essential that there be sufficient power in its test. If there were not enough power, the researcher might mistakenly conclude that the data are independent when they are not. Table 2 presents the power tables for the test of the intraclass correlation for an alpha of .05, two-tailed. We used a method described by Koele (1982) to estimate power. Three factors are varied in the table: group size, overall sample size, and the degree of nonindependence. It is assumed that there is a single nested independent variable with two levels.

Not surprisingly, power increases as the intraclass correlation and sample size increase. Generally, there is more power when group size is larger, unless there are very few groups ($N = 40$ and $n = 10$). We see that especially when the intraclass correlation is not large and total sample size and the group size are small, power is very low. For example, with twenty groups of five persons and an intraclass of .15, the probability of making a Type II error is .56. Because of this low power, it is advisable to raise alpha to .2 in the test of the intraclass correlation (Myers, 1979). We return to the issue of power in the test of the intraclass correlation in the "General Guidelines" section.

Effect of Nonindependence on Tests of the Independent Variable To what degree does ignoring nonindependence bias tests used to determine whether the treatment (factor A) has a statistically significant effect? That is, if there are group effects but person (ignoring group) is used as the unit of analysis, does the p value associated with the obtained F statistic truly represent the likelihood of obtaining that F if the null hypothesis were true and there were no effects due to the independent variable?

To determine the effect of nonindependence on the effective alpha for the three types of independent variables, a three-step procedure developed by Kenny (1995) is used. First, the critical value for the F test with person as unit is determined for the degrees of freedom. That critical value is divided by a bias factor (denoted as B in Table 3) to produce an adjusted F that is then used to determine the adjusted critical value. The bias factor B depends on the type of independent variable (i.e., nested, crossed, or mixed) and the size of the intraclass correlation. Finally, the degrees of freedom for the adjusted F test are reduced given the type of design and the size of the intraclass (denoted as df' in Table 3). The p value associated with the adjusted F and the adjusted degrees of freedom gives the effective alpha for the test.

The first row of formulas in Table 3 presents the formulas for the bias factor and the corrected degrees of freedom for a nested independent variable. For instance, if the total sample size is 100, the nonindependent observations are pairs of observations ($n = 2$), and the intraclass correlation of ρ is .5, then the bias factor which divides F is 1.52, and the effective degrees of freedom are 78.08, not 98. So the F test is inflated by about 50 percent, and the real degrees of freedom are about 20 less. Formulas for other designs are also presented in Table 3 and are used later in this section.

Table 4, using the formulas contained in Table 3, presents the value of the effective alpha when person is used

TABLE 2

Power (Times 100) of the Test of the Intraclass Correlation (ρ) with Two-tailed Alpha of .05 for the Nested Design

As a Function of the Size of the Correlation, Group Size (n), and Total Sample Size (N)

	N								
	40			**100**			**200**		
	n			***n***			***n***		
ρ	**2**	**5**	**10**	**2**	**5**	**10**	**2**	**5**	**10**
−.05	5	5	5	6	9	15	8	16	37
.05	6	8	10	6	11	16	8	17	27
.15	10	21	25	18	44	57	32	72	85
.25	19	40	40	42	78	82	71	97	98
.35	33	60	54	71	94	94	95	100	100
.45	53	76	65	92	99	98	100	100	100

Note: There are two treatment groups.

TABLE 3

Corrections to the *F* test (*B*) and the Degrees of Freedom (*df'*) for Error

Nested

$$B = \frac{1 + \rho(n-1)}{1 - \rho(n-1)/(N/2-1)}$$

$$df' = \frac{[N-2-2\rho(n-1)]^2}{N-2+\rho(n-1)\{\rho[N-2(n-1)]-4\}}$$

Crossed

$$B = \frac{1-\rho}{1-\rho(n-2)/(N-2)}$$

$$df' = \frac{[N-2-\rho(n-2)]^2}{N-2+\rho^2[n-2+(n-1)(N-n)]-2\rho(n-2)}$$

Crossed with Interaction

$$\rho_1 = \frac{\rho_G}{1+\rho_{GxA}/(1-\rho_{GxA})}$$

$$\rho_2 = \frac{\rho_{GxA}}{1+\rho_G/(1-\rho_G)}$$

$$B = \frac{(N-2)[1-\rho_1+\rho_2(n/2-1)]}{N-2-\rho_1(n-2)-\rho_2(n/2-1)/2}$$

$a = (N-n)n/2$

$b = (n-2)(N-n+2)/2$

$c = 2(n-2)$

$$df' = \frac{[N-2-\rho_1(n-2)-\rho_2(n/2-1)/2)]^2}{N-2+a\rho_1^2+b(\rho_1+\rho_2)^2-c(\rho_1+\rho_2)}$$

Mixed (*n* = 2)

$$B = 1 + \frac{N\rho_x\rho_y}{N-(1+\rho_x)(1+\rho_y)}$$

$$df' = \frac{d}{e+f+g}$$

where

$d = (N-2)[N-(1+\rho_y)(1+\rho_x)]^2$

$e = \rho_x^2 + [\rho_y^2(N^2-3N+1)]$

$f = (N-1)[N-2(1+\rho_y+\rho_x)+1]$

$g = \rho_y\rho_x[\rho_y\rho_x(N+1)+2\rho_x-2\rho_y(N-1)-2(N-2)]$

Note: N is the total number of observations in the study, *n* the group size (and so the number of groups is *N/n*), and ρ the degree of nonindependence. For the mixed design, *n* equals 2, ρ_x represents nonindependence of the independent variable, and ρ_y represents nonindependence of the dependent variable. For all cases, there are two treatment groups.

as the unit of analysis even though there are group effects. Varied in the table are the total sample size *N*, the degree of nonindependence as measured by the intraclass correlation ρ, and the number of persons per group *n*. In this table there are always two treatment conditions; the total sample size is 40, 100, or 200; and the group size is 2, 5, or 10. Note that the total number of groups in a study is the total sample size divided by the group size. If alpha is greater than .05, then the statistical test is said to be too *liberal* and the null hypothesis is rejected too often. If alpha is less than .05, the test is said to be too *conservative* and the test has artificially low power.

As seen in Table 4, the degree of distortion in alpha depends on several factors. Looking first at the total sample size, it appears that *N* has virtually no impact on alpha. Only with large intraclass correlations can any effect be seen; with the larger sample size, there is slightly less distortion in alpha. Because the number of groups equals *N/n*, that factor too has virtually no effect on alpha.

The intraclass correlation is an important factor in alpha distortion. If the intraclass correlation is negative, the test is too conservative. That is, the null hypothesis is not rejected as often as it should be. As the value of the intraclass correlation becomes more positive, however, the test becomes increasingly liberal. For large values of the intraclass correlation, the null hypothesis is rejected much more often than it should be. Group size is a second important factor in determining the bias. With larger group sizes and a fixed total sample size, the larger the group size, the greater the alpha distortion.

To summarize, alpha is affected by the size of the intraclass and the group size but not affected very much by the total sample size or the number of groups. In essence, when the intraclass correlation and group size are both large but person is used as unit, the *p* values are grossly inflated. Fortunately, it seems likely that as group size increases, the intraclass declines (see Latané, 1981). As shown in Table 2, the power of the test of the intraclass correlation is affected by *N*. A small *N* study would have very little power in the test of the intraclass correlation and the intraclass correlation may be fairly large but not significant. For such a study there might be substantial distortion in the effective alpha value.

When the intraclass correlation is small and there are few persons per group, the distortion in alpha is fairly small. For instance, when the intraclass correlation is .05 and there are just two persons per group, using person as the unit of analysis results in a slightly too liberal test, the alpha being only .06.

Power Clearly ignoring nontrivial levels of nonindependence can seriously distort alpha. What then are the consequences if, because of nonindependence, group instead of person is used as the unit of analysis when testing for differences in the effects of strategy type? The second part of

TABLE 5

Power (Times 100) with Group as the Unit of Analysis for the Nested Design with Two Treatment Groups and 100 Persons

	n															
	1*	2					5					10				
	ρ_G	ρ_G					ρ_G					ρ_G				
ES†	—	.0	.1	.2	.3	.4	.0	.1	.2	.3	.4	.0	.1	.2	.3	.4
0.2	17	15	14	14	13	12	15	12	10	9	9	14	9	8	7	7
0.5	70	69	65	61	57	54	66	51	41	34	30	59	34	25	20	17
0.8	98	99	98	96	95	93	98	91	82	73	65	95	73	55	44	37

*Baseline condition; person as unit with independence.
†Effect size (Cohen's *d*).

in power. When group becomes the unit, sample size declines by a factor of *n* (the number of persons per group), but the adjusted effect size increases by that same factor. These two factors nearly exactly offset each other. Given a sufficient number of groups (about twenty or more groups for the entire study), power is virtually unaffected when group is the unit of analysis and the intraclass correlation is zero. Thus, mistakenly making group the unit of analysis (when person should be) has little effect on power, at least when there are a sufficient number of groups. If the intraclass correlation is negative, the power increases when group is used as the unit of analysis. However, as the intraclass correlation increases, power does decline, especially when there are many persons per group. So there is less power when group scores are not independent.

When a researcher is faced with low power in a group study, there are, in principle, two ways that power can be enhanced (assuming a positive ρ). Either the number of groups (*N/n*) can be increased or the number of persons per group (*n*) can be increased. The latter strategy is available when the groups are large in size (e.g., classrooms or organizations) and the researcher samples a larger subset of the members. Consider the following example: a researcher has two treatment groups, within each group there are five classrooms, and from each classroom five students are sampled. For a large effect size ($d = 0.8$) and an intraclass correlation of .25, the power of the test is .42. Power can be increased by doubling the number of students per classroom to ten, and now the power is .50. But if instead the total number of classrooms in the study is doubled from ten to twenty and each group still has five students, the power climbs to .76. Both studies have 100 students (one with ten groups or classrooms each with ten students and the other with twenty groups each with five persons), but the second has considerably more power. The lesson to be learned is that it is generally better to add more groups to the study than it is to increase group size when the intraclass correlation is nontrivial.

General Guidelines for Nested Independent Variables If there is nonindependence, then group not person must be used as the unit of analysis. So in principle, the researcher should first evaluate whether there is nonindependence. If there is nonindependence, then group may be the unit of analysis; and if there is independence, then person may be the unit of analysis. To determine if there is nonindependence, the intraclass correlation is estimated and tested for statistical significance. There is one major sticking point with this procedure: the test of nonindependence may be very low in power. The usual standard for "sufficient power" is having an 80 percent chance of rejecting the null hypothesis (Cohen, 1988).

Given nonindependence and using person as the unit of analysis, there is bias in the test of the treatment effect. What is a reasonable value for the largest possible bias that researchers would accept? Of course, we would wish that there was no bias in alpha, but we are willing to tolerate small distortions in alpha for trivial levels of nonindependence. Because it has become fairly routine to treat *p* values between .05 and .10 as marginally significant, it would seem that .10 is the largest possible bias that most social psychologists would be willing to tolerate. Therefore, we define *consequential nonindependence* as the level of the intraclass correlation that occurs when person is inappropriately treated as the unit of analysis, and, as a result, the test of the independent variable is biased such that the nominal value alpha of .05 actually corresponds to an alpha of .10.[3]

Table 6 presents the minimum number of groups needed to detect consequential nonindependence. The rows in this table are group sizes and the columns are the alphas used to test the intraclass. The first column uses the standard alpha of .05 for the test of the intraclass correlation, and the second column uses the more liberal value of alpha of .20 which has been recommended by some authors (Myers, 1979).[4]

We see in Table 6 that thirty-six dyads are required to

TABLE 6

Minimum Number of Groups Needed to Have at Least .80 Power to Detect Consequential, Positive Nonindependence (Effective Alpha of .10)

As Function of Group Size and the One-tailed Alpha Used to Test the Intraclass Correlation

	Number of Groups (N/n) One-tailed Alpha to Test ρ	
Group Size (n)	**.05**	**.20**
2	36	18
3	56	28
4	66	34
5	72	36
6	76	38
7	80	40
8	82	40
9	82	40
10	84	42

achieve the standard of 80 percent power. Thus, to have an 80 percent chance of detecting nonindependence that biases the test of the independent variable to an effective alpha of .10, a minimum of thirty-six dyads are required. For groups of size eight, eighty-two groups (656 persons!) would be needed. The large number of groups needed when group size is large results from the fact that for large-sized groups, a very small level of nonindependence can create serious distortion of p values (see Table 4). Fewer groups are needed if the test of the intraclass is made more liberal, but the number of groups required is still substantial for large groups.

The implication of Table 6 is that the general advice given above is practical for studies with dyads and maybe triads, but it is not useful for studies in which groups are composed of four or more group members. For these studies, there is ordinarily insufficient power to test for consequential nonindependence.

If there are not enough groups to have a powerful test of the intraclass, group should be the unit of analysis. Unless the experimental procedure or previous research strongly indicates that the data are independent, group research requires using group as the unit. Fortunately, as shown in Table 5, there is surprisingly little loss of power in using group as unit when there is nonindependence.

Some researchers may be unwilling to use group as unit, or in some cases there may be so few groups per condition that there may be too little power to make group the unit. If person is the unit (something we do not recommend!), then *S/G/A* should be used as the error term to test for the effect of the independent variable, and the group variance (*G/A*) should still be removed. This approach of treating person as the unit has the advantage of removing the group effect from the error term, but it has the disadvantage of limiting the conclusions to the specific groups studied. Effectively, group is treated as a fixed, not a random, factor.

It might be argued that group should be treated as a random factor only when the groups are randomly sampled from the population of groups. However, persons are hardly ever randomly sampled, yet researchers routinely treat them as if they were random. Although sampling considerations are important in statistical decision making, it does not seem justifiable to insist on random sampling of groups and not to insist on random sampling of persons.

If person is used as the unit and the test of the independent variable is statistically significant, the fail-safe correlation (Kenny, 1995) can be computed. This correlation estimates how large nonindependence would have to be to render what is a statistically significant result no longer significant. An approximation to the fail-safe r is the following:

$$\frac{F - F_c}{F_c(n-1) + F(n-1)/(m-1)}$$

where F is the test statistic for A, F_C is the critical value for that test, m is the number of persons at each level of the independent variable, and n is the number of persons per group. If the fail-safe r is implausibly high, then nonindependence is not a plausible rival explanation of a significant result. We should make it clear that we do not recommend using person as unit. But if the researcher does not follow this advice, computing a fail-safe r would be advisable.

Crossed Independent Variable

Nested independent variables are much more frequently used in group research than crossed independent variables. In a crossed design, some members of each group are in one treatment condition whereas other members of the same group are in the second treatment condition. Thus, in this design, condition and group are crossed.

Consider another hypothetical study: a researcher studies the effect of gender in group communication and forms twenty-five groups. In each group there are two men and two women, and the total sample size is 100. Table 7 presents the ANOVA table for the study. Included in this table is the main effect of the experimental factor, a person's gender, denoted as factor A. This effect measures whether men or women talk more. The second factor in the table is the main effect of Group (G) which measures the extent to

TABLE 7

ANOVA Source Table for the Crossed Design Example with Twentv-Five Groups of Four Persons

Source	*SS*	*df*	*MS*
Between Group			
Group (G)	120	24	5
Within Group			
Gender (A)	25	1	25
GxA	72	24	3
Person (S/GxA)	100	50	2

$$\rho_{GxA} = \frac{MS_{GxA} - MS_{S/GxA}}{MS_{GxA} + [n/q - 1]MS_{S/GxA}}$$

(where n is the number of persons per group and q the number of levels of A)

$$\rho = \frac{3-2}{3+[4/2-1]2} = .20$$

$$F(24{,}50) = \frac{MS_{GxA}}{MS_{S/GxA}} = 1.50, p = .226, \text{two-tailed}$$

which people in some groups talk more than people in other groups. The next source of variation is the interaction between group and gender (GxA) which measures the extent to which, in some groups, gender differences are larger than in other groups. The final source of variation is person within the group X gender interaction (S/GxA) which measures the extent to which some persons talk more than others controlling for both group and gender. In general, for a crossed variable, there are n persons in each group and n/q persons at each of q levels of the independent variable. Note that if the researcher had studied opposite-gender dyads (groups of size two), the person within group by gender term could not be estimated. That is, within each dyad there would be only one male and one female and so variation within gender cannot be computed.

One key advantage of the crossed design over the nested design is that the group main effect and the group X condition interaction can be separated. In the nested design the group and the group X condition interaction are combined in the G/A term, and variance due to both is contained in the mean square for treatment. However, in the crossed design the condition effect contains only the group X condition interaction variance and not the group main effect variance. Thus, in the crossed design the effect of the independent variable is, at least in principle, estimated with greater precision and therefore tested with greater power than in the nested design.

For the crossed design, treating group as the unit of analysis involves testing the effect of the independent variable A, using the group X treatment interaction mean square or MS_{GxA}. For the fictitious study presented in Table 7, that test is $F(1{,}24) = 8.33$. If person is the unit of analysis, there are three possible ways to test the A effect. First, S/GxA could be used as the error term such that $F(1{,}50) = 12.50$. Alternatively, the group X treatment interaction (GxA) could be pooled with S/GxA to yield a pooled error term of 2.32 ([72 + 100]/[24 + 50]). With this error term, the test of A yields $F(1{,}74) = 10.76$. Finally, both the effects of group and its interaction with gender can be pooled with $MS_{S/GxA}$, and the resulting mean square error would equal 2.98 ([120 + 72 + 100]/[24 + 24 + 50]). The test of the independent variable would be $F(1{,}98) = 8.39$. In the crossed design there are four alternative error terms. In the example, the mean square error term ranges from 2.00 to 3.00, the degrees of freedom from 24 to 98, and the F from 8.33 to 12.50.

Measuring the Group Main Effect and the Condition by Group Intraclass Correlations If person is treated as the unit of analysis in the crossed design there are two potential sources of nonindependence in the data: the group main effect and the group X condition interaction. The presence of either of these sources of variance results in nonindependence and invalidates the use of person as the unit of analysis. The intraclass correlation on the outcome measure for the group effect, or ρ_G, can be measured and tested for statistical significance as before with the nested design. The ingredients are the mean square for group which equals 5 in the example (see Table 7), the mean square for person within the group X condition interaction, which equals 2, and the total number of persons per group or 4. The value of ρ_G for the hypothetical example is $[5 - 2]/[5 + (4 - 1)2] = .27$.

The bottom of Table 7 shows how the intraclass correlation for the interaction, or ρ_{GxA}, can be assessed. The ingredients for the formula are the mean square for the interaction, the mean square for persons within this interaction, and the number of persons within each group and condition, two for the example. Like the group intraclass correlation, the interaction intraclass correlation can be tested by an F test. The lower limit of ρ_{GxA} is $-1/[n/q - 1]$ where q is the number of conditions.[5]

If the intraclass correlation for the interaction is positive, it means that within a group, the two women's levels of talking are more similar to one another than to the two men's; and correspondingly, the two men's levels of talking are more similar to each other than to the women's. Alternatively, the correlation implies that the gender difference in talking varies from group to group.

As stated earlier, if the group size equals the number of levels of the independent variable, it is not possible to separate variation due to S/GxA from variation due to GxA. Thus the group X condition interaction (GxA) cannot be tested and the intraclass correlations for *GxA* cannot be estimated. It is still possible to estimate the value of ρ_G by substituting MS_{GxA} for $MS_{S/GxA}$.

There has been very little systematic investigation of the size of the *GxA* interaction. However, it seems reasonable to expect that variance due to this interaction is fairly small. Usually, the group *X* condition intraclass correlation is smaller than the group intraclass ($\rho_{GxA} < \rho_G$); however, there are certain to be exceptions to this rule.

The power of the test for ρ_G is comparable to that for the nested design (see Table 2). The power of the test of ρ_{GxA} is likely even lower than the test of ρ_G for two reasons. First, ρ_{GxA} is usually smaller than ρ_G (see above), and second, for ρ_{GxA} the effective sample size is not *n* but rather *n/q* and smaller group sizes result in lower power (see Table 2).

Effect of Nonindependence on Tests of the Independent Variable What are the consequences of ignoring nonindependence by treating person as the unit of analysis in the crossed design? Consider an example in which there are 100 persons and two conditions. Assume first that the group *X* condition interaction ρ_{GxA} is zero. When the group intraclass correlation ρ_G is negative, the test of factor *A* is slightly too liberal. However, in the much more likely case of a positive intraclass, the test is too conservative and therefore artificially low in power.

When the intraclass correlation for group is positive, the design is akin to a repeated measures design in the sense that each group has persons in each condition. If group is treated as the unit of analysis, variance due to group is subtracted from the error term that is used to test the treatment effect. Thus, treating group as the unit increases power in the same way as a within-subjects design has more power than a between-subjects design.

What, then, is the effect of treating person as the unit of analysis when there is nonindependence due to both the group and the group *X* condition interaction? As can be seen in Table 8, for dyads ($n = 2$) the interaction intraclass correlation is irrelevant. If group size is greater than two, the test of factor *A* becomes somewhat more liberal as the intraclass correlation for the interaction increases and as the intraclass correlation for group decreases. Also, as the number of groups declines (and hence the number of persons per group increases) the alpha inflation increases. Interestingly, when there are four persons per group and the intraclass correlation for groups equals the intraclass correlation for interaction, the two sources of bias virtually cancel each other. In addition, note the parallel between the crossed design with interaction effects and weak group effects and the nested design with group effects. For both designs, as the degree of nonindependence increases, the value of the effective alpha also increases.

Power The determination of power of the test of a crossed independent variable depends on the effect size, the total sample size (*N*), the group size (*n*), the intraclass correlation for group, and the intraclass correlation for the group *X* condition interaction. Using the Severo and Zelen (1960) approximation, Table 9 presents the power for a moderate effect size ($d = 0.5$); total sample sizes of 40 and 80; groups of size 2, 4, and 10; and intraclass correlations of −.10, .00, .20, and .40. The values in the table are based on the assumption that group is treated as the unit of analysis and so the *A* effect is tested by the *GxA* interaction. As a reference point, the power with person as unit and independence is .34 for $N = 40$ and .60 for $N = 80$.

The first thing to note in the table is that power increases as the intraclass correlation for group or ρ_G increases. Because variance due to group is removed, power is increased. However, power declines as ρ_{GxA} increases if the group size is larger than two. So for groups larger than two, the increase in power due to removing group variance can be lost if the variance due to group *X* condition interaction is large. Also, increasing group size lowers power, but this loss of power is most pronounced as ρ_{GxA} increases.

General Guidelines for Crossed Independent Variables The best general advice to give for a crossed independent variable is to treat group as the unit of analysis. This would be accomplished by evaluating the effect of the independent variable using the group *X* condition interaction as the error term. This approach results in the removal of the group main effect which generally increases the power of the test. There are several reasons for this advice. First, very often crossed group designs are dyadic and so each group has just one person in each condition. In this case, the group *X* condition interaction is the only available error term. Second, even when each group has more than one person in each condition, usually there is more power in the test of a crossed independent variable when group, not person, is the unit of analysis.

When each group has more than one person in each condition, the two-stage strategy discussed for nested independent variables is an option. First, one estimates and tests the group *X* condition interaction using the $MS_{S/GxA}$ as the error term. We recommend as before using a liberal alpha of .20. If the interaction is significant, then group must be treated as the unit of analysis if effects are to be generalized beyond the specific groups studied.

If the test of group *X* condition is not significant, then person can be treated as the unit of analysis. Using person as the unit involves pooling the group *X* condition sums of squares (SS_{GxA}) with the sum of squares for persons within the group *X* condition interaction ($SS_{S/GxA}$). Similarly, the degrees of freedom from these two effects should be pooled. Pooling these two sources of variation should provide a more powerful test of a crossed independent variable than would occur if the $MS_{S/GxA}$ alone were used as the error term.

The major drawback of this two-stage procedure is that

TABLE 8

Effect of Group (ρ_G) and Group by Treatment (ρ_{GxA}) Correlations and Group (n) Size on Alpha (Times 100) with Person as the Unit of Analysis

	n														
	2					4					10				
ρ_{GxA}	ρ_G					ρ_G					ρ_G				
	–.1	.1	.2	.3	.4	–.1	.1	.2	.3	.4	–.1	.1	.2	.3	.4
–.1	6	4	3	2	1	5	3	2	1	1	2	1	0	0	0
.1	6	4	3	2	1	7	5	4	3	2	11	9	8	7	6
.2	6	4	3	2	1	9	6	5	4	3	16	14	13	12	11
.3	6	4	3	2	1	10	7	6	5	4	20	18	17	17	16
.4	6	4	3	2	1	11	9	8	6	5	24	22	22	21	20

Note: There are 100 persons and two treatment groups. The number of groups varies from fifty, twenty-five, to ten groups. The intraclass correlations are adjusted as in Table 3.

the test of the group X condition interaction may often have very low power. Unless it can be established that there is sufficient power, we feel the safest course of action is to use the group X condition interaction as the error term in the test of the independent variable.

Sometimes because of poor design, there may be too few groups to make group the unit of analysis. In this instance, the researcher may be forced to treat group as a fixed effect and person as the unit of analysis. However, the conclusions from such an analysis would refer to the specific groups that were sampled resulting in little generalizability.

Mixed Independent Variable

A mixed variable varies both between and within groups. For example, if one were studying romantic relationships (so n equals 2) and included gay as well as heterosexual couples, gender would be a mixed variable. A second example of a mixed variable is intelligence level in a study in

TABLE 9

Power (Times 100) for the Crossed Design with a Medium Effect Size (d = 0.5)

		n											
		2				4				10			
		ρ_G				ρ_G				ρ_G			
N	ρ_{GxA}	–.1	.0	.2	.4	–.1	.0	.2	.4	–.1	.0	.2	.4
40	–.1	28	31	38	48	28	31	38	51	26	30	41	66
	.0	28	31	38	48	26	28	34	44	19	20	24	30
	.2	28	31	38	48	23	24	28	34	13	13	15	16
	.4	28	31	38	48	20	21	24	28	11	11	11	12
80	–.1	54	58	69	82	56	61	72	87	63	70	87	99
	.0	54	58	69	82	52	56	66	79	45	48	58	70
	.2	54	58	69	82	45	48	56	66	28	29	32	36
	.4	54	58	69	82	40	42	48	56	21	22	23	25

Note: There are either forty or eighty persons and two conditions. Group is the unit of analysis. The intraclass correlations are adjusted as in Table 3.

which persons are randomly assigned to groups of four persons. Intelligence is a mixed variable because within groups some persons would be more intelligent than others, and some groups would have a higher average intelligence than others.

One way to determine whether an independent variable is mixed, nested, or crossed is to compute its intraclass correlation. To compute that correlation for the independent variable, denoted as ρ_x in this section, the independent variable is treated as a dependent variable. If the independent variable is nested, the intraclass correlation equals one; if the independent variable is crossed, the intraclass equals its lower limit; and if the independent variable is mixed, the intraclass correlation is not at either limit. The intraclass correlation for the outcome variable is denoted as ρ_y.

In our discussion of mixed variables we focus on three issues. First, we consider the effect of using person, rather than group, as the unit of analysis thereby ignoring nonindependence. Second, we discuss how mixed designs, unlike nested or crossed designs, allow the estimation of partner effects: the degree to which a person's level of the independent variable affects the dependent variable scores of other group members. Third, we describe the statistical analysis of mixed independent variables.

Effect on *p* Values As noted, a mixed variable's intraclass correlation is less than one and greater than $-1/(n - 1)$ where n is the group size. When the independent variable is not a manipulated variable, it is often mixed. For instance, in an investigation of the effect of attraction toward the group on work performance in groups, it is likely that attraction varies both between persons (some persons are more positive about the group than others) and between groups (some groups on average are more positive than others). Thus, attraction is likely to be a mixed independent variable with an intraclass correlation that is positive, but not perfect.

Kenny (1995) presents the details on how to compute the intraclass correlation when a variable is assumed to be caused by a mixed variable. The bias to the F test is quite complicated for the mixed case. Fortunately, for dyads the bias to F is relatively simple (Kenny, 1995) and is presented in Table 3. However, the adjustment to the degrees of freedom is very complicated, even for dyads.

Table 10 presents the bias in the test of the effect of a mixed independent variable when group size is limited to dyads. In this table, there are assumed to be fifty dyads (100 persons). When ρ_x and ρ_y have the same sign, the F test is inflated and so the test is too liberal. Importantly, there is not as much bias in the test when the independent variable is mixed as when it is nested or crossed. Because a mixed variable is between a nested and a crossed variable and because the effect of nonindependence is the opposite for nested and crossed designs, mixed independent variables create relatively weak effects due to nonindependence. In fact, if ρ_x and ρ_y are less than or equal to .3, alpha never exceeds .061. Generally, for mixed independent variables there is less distortion of alpha than there is for nested or crossed variables.

TABLE 10

Effect on Alpha of the Intraclass Correlation of the Independent (ρ_x) and Dependent Variable (ρ_y) for the Mixed Design for 100 Persons and 50 Dyads

	ρ_x						
ρ_y	**–.5**	**–.3**	**–.1**	**.0**	**.1**	**.3**	**.5**
–.5	.080	.068	.056	.051	.045	.034	.024
–.3	.068	.061	.054	.051	.047	.041	.034
–.1	.056	.053	.051	.050	.049	.047	.044
.0	.050	.050	.050	.050	.050	.050	.050
.1	.044	.047	.049	.050	.051	.053	.056
.3	.034	.040	.047	.050	.054	.061	.068
.5	.024	.034	.045	.051	.056	.068	.080

Note: The tabled values are the actual p values when the test statistic has a nominal p value of 0.05.

To determine the approximate power for a mixed independent variable, we can extrapolate from the power analyses that were done when the independent variable is nested or crossed. If the product of the two intraclasses $\rho_x\rho_y$ is positive, power tends to decline when dyad, not person, is the unit of analysis. Alternatively, if $\rho_x\rho_y$ is negative, power is usually increased by using dyad as unit. If ρ_x is near zero, there is little effect on power. So for small values of ρ_x, both power and p values are not much affected.

Partner Effects Nested, crossed, and mixed independent variables all allow for the examination of the degree to which a person's score on the independent variable affects that person's score on the response variable. With a mixed independent variable, however, it is possible to estimate a second effect that cannot be estimated with either a nested or a crossed independent variable. This second effect measures the degree to which one person's score on the independent variable affects the responses of the other persons in the group. To distinguish these two types of effects, the former is called an *actor* effect and the latter a *partner* effect (Kashy & Kenny, 1997; Kenny, 1996b). As an example, consider again the effect of attraction toward the group on productivity. The actor effect measures whether persons who are highly attracted to the group are more productive. The partner effect measures whether being in a group with a person who is highly attracted to the group results in the other group members being more productive. Kashy and Kenny (1997) present a detailed discussion of what they call the *Actor-Partner Interdependence Model.*

Although the sizes of actor and partner effects are in principle independent, there are two special cases that are particularly relevant to the study of group behavior in social psychology. In the first case, which we will call *group orientation,* the actor and partner effects are approximately equal. In the second case, called *social comparison,* the actor and partner effects are equal in magnitude but opposite in sign, such that the actor effect is usually positive and the partner effect is negative. To illustrate these two types of effects, consider the effect of physical attractiveness on outcomes. If the model is group orientation, then a person's physical attractiveness leads to positive outcomes for the person and his or her partners. If the model is social comparison, then physical attractiveness leads to better outcomes for the self but reduced outcomes for the partner. Tesser's (1988) self-evaluation maintenance model explicitly considers these two types of effects.

Although the concept of partner effects may seem to be new, it has been used extensively in work in social psychology, particularly in small group theory and game theory. Consider first its use in small group theory. Partner effects provide an empirical method of gauging whether the group is one in which people see others as part of self or as different from self. If there are group-norm effects, then the success of the other leads to feelings of happiness, as much happiness as one's own success; it really does not matter who succeeds. If there are social comparison effects, then the success of the other leads to unhappiness.

Partner effects can also be used to define the classic prisoner's dilemma (PD) game. In this game, two people are given two choices. Each of these outcomes depends on their joint choice. If behavior choice is the independent variable (cooperative versus competitive) and outcome the dependent variable, then the essence of PD is that the actor effect and partner effect have different signs: actions that lead to better outcomes for the self have negative consequences for the partner. In fact, the defining feature of PD is that partner effects are larger than actor effects: a person's choice affects the partner's outcome more than that person's. Kelley and Thibaut (1978) decompose outcomes in two-person games into actor and partner effects, but they use different terms for the two effects.

Statistical Analysis Perhaps the first explicit recognition of the analysis difficulties raised by a mixed design, as well as a realization that partner effects can be estimated, is work by Kraemer and Jacklin (1979). Their approach though ground-breaking is limited: the independent variable can be only a dichotomy; there can be no additional independent variables; and the tests are large sample tests. Kenny (1996b) has developed two generalizations of the Kraemer and Jacklin (1979) approach. The first generalization involves computation of two regression equations. In one regression, the group average of the independent variable is used to predict the group average of the response variable. In the second regression, the deviations from the group mean for both the independent and dependent variables are used in a regression equation. The regression coefficients from these two analyses are then pooled (Kashy & Kenny, 1997; Kenny, 1996b) to estimate the actor and partner effects. The second generalization involves the use of structural equation modeling (e.g., LISREL) and is described in detail in Kenny (1996b). These generalizations presume that group sizes are equal. A more general procedure for analyzing mixed designs, one that does not presume equal group sizes, is discussed in the multilevel modeling section of this chapter.

General Guidelines for the Analysis of Mixed Variables An examination of Table 10 indicates that the value of the intraclass correlations must be very large to have consequential effects on the significance testing of the independent variable. If the intraclass correlation is not larger than .5, it is safe to use person as the unit of analysis, at least for dyads. We need further study about the effect of nonindependence and mixed variables for groups. The more important issue with mixed variables is to estimate partner effects because partner effects capture the truly interpersonal nature of group interaction. The analysis of data from mixed designs poses special difficulties that cannot be handled within a traditional ANOVA framework.

Summary

Most group research contains not one independent variable but multiple independent variables. Likely, some variables are nested, others are crossed, and some are mixed. As we have seen, a given level of nonindependence has very different effects for the different types of independent variables. For instance, if the intraclass correlation of the outcome variable were .45 and person is the unit, for a nested variable the F test would be too large, for a crossed variable it would be too small, and for the mixed variable there may be little or no effect.

Generally the safest course of action is to make group the unit of analysis and so it is then necessary to collect data from a sufficient number of groups. Although there may be some loss of power (perhaps not nearly as much as might be thought), treating group as the unit avoids many of the problems detailed in this section.

MULTILEVEL MODELS

In the previous section, ANOVA provided the framework for analyzing data from nested and crossed independent variables. As noted, ANOVA models cannot be used to analyze data from mixed variables; these models are furthet limited by the assumptions that group sizes are equal and the independent variables are categorical. There is an alternative to the ANOVA approach that can be used with any of the three types of independent variables which allows both continuous independent variables and unequal group sizes. This data-analytic approach has many names: multilevel models, hierarchical linear models, mixed model ANOVA, and random regression estimation.

The defining feature of multilevel data is that there is a hierarchy of observations. The lower level is the level at which the outcome variable is measured and is nested or grouped within an upper-level unit. In group research, the lower-level unit is person, and the upper-level unit is group. Applications of multilevel modeling are not, however, limited to group research. Most especially, in repeated-measures research (e.g., diary studies), observation or time point can be treated as the lower-level unit, and person can be treated as the upper-level unit. There is no requirement that persons have the same number of data points, as there is in repeated-measures ANOVA.

The Basic Data Structure

As an example of the basic data structure, we consider a fictitious study examining the effects of leadership style, authoritarianism, and satisfaction of group members. The participants in this "study" are military recruits in twenty-three platoons. Platoons range in size from five members to fifteen members, and recruits are randomly assigned to platoons with either a democratic or an autocratic leader. At the beginning of the study, all the recruits are pretested on authoritarianism. After six weeks of boot camp, the recruits rate their level of satisfaction with their platoon.

This study can be used to investigate several questions concerning leadership style, authoritarianism, and group satisfaction. First, it can test whether recruits generally are more satisfied with autocratic or democratic leadership styles. This first question concerns the effects of a nested independent variable, and ANOVA could be used to analyze such data if the leadership effect were the sole question of interest. The second question that can be addressed by this study is whether authoritarianism predicts satisfaction. Authoritarianism is a mixed independent variable because some recruits in a platoon are more authoritarian than others and some platoons score higher on average in authoritarianism than others. The third question that this study can address, using a multilevel approach, is the interaction between leadership style and authoritarianism: are recruits who are higher in authoritarianism more satisfied with autocratic leaders?

In this example, recruit is the lower-level unit and platoon is the upper-level unit. Authoritarianism is a lower-level predictor variable which we symbolize as X. Note that X can be either continuous or categorical (categorical Xs are dummy coded). When X is continuous, as is the case for authoritarianism, it should be centered (Aiken & West, 1991) so that the intercepts are more interpretable. To center the variable, we subtracted the grand mean of authoritarianism (i.e., 5.53) from each score in this artificial data set. Leadership style is an upper-level predictor variable and is denoted Z. Like X, Z can be either continuous or categorical. In the present example, leadership style is effect coded: $Z = -1$ for the democratic style and $Z = 1$ for the autocratic style. Again, if Z were continuous, it should be centered. Finally, the outcome variable, satisfaction with the platoon, is measured on a seven-point scale at the lower level (recruit) and is symbolized as Y. Table 11 presents selected observations from this fictitious example data set.

Note that we are not allowing for partner effects (see the discussion of mixed variables in the previous section). For the example, the partner effect would refer to the effect of the authoritarianism of the other platoon members on the recruit's satisfaction. Had we wished to allow for such effects, we would create an additional X variable: the mean authoritarianism of those in the platoon besides the recruit (Kenny, 1996a).

Consistent with the conclusions drawn in the unit of analysis section, in multilevel modeling the upper-level unit, platoon in the example, is treated as the fundamental unit of analysis. The basic analysis is a two-step procedure in which an analysis is performed within each upper-level unit (platoon) and then the results of all these analyses are pooled. That is, in the first step the relationship between X and Y is estimated for each upper-level unit. In the example data set, the first step of the analysis estimates the relationship between authoritarianism and satisfaction separately for each platoon. In the second step, the results from the step one analyses are pooled across the upper-level units and the effects of the upper-level predictor variable, leadership condition, are assessed.

Unweighted Regression

It is usually advisable to relate X and Y by a regression analysis. So for each upper-level unit, Y is regressed on X. In the example, for platoon i with recruit j, the model is Equation 1 where platoon i has its own intercept b_{0i} and slope b_{1i}:

$$Y_{ij} = b_{0i} + b_{1i}X_{ij} + e_{ij} \qquad (1)$$

This analysis approach presumes that there are at least three observations for each group and that both X and Y vary for each group. There are k groups, n_i observations for group i (so there are n_i recruits in platoon i), and N or Σn_i total observations. The number of observations per group need not be equal. These k regressions are called the *first-step* regressions.

TABLE 11

Selected Observations from the Leadership and Authoritarianism Hypothetical Data Set

Platoon Number	Leadership Condition	Recruit Number	Authoritarianism	Satisfaction
2	−1	7	−0.53	4
2	−1	8	−1.53	3
2	−1	9	−2.53	4
2	−1	10	2.47	5
2	−1	11	2.47	4
2	−1	12	0.47	3
2	−1	13	−0.53	6
4	1	22	0.47	6
4	1	23	1.47	4
4	1	24	1.47	5
4	1	25	3.47	6
4	1	26	2.47	5

For the second-step analysis, the regression coefficients estimated in the first-step regressions (see Equation 1) are assumed to be a function of the upper-level predictor variable Z:

$$b_{0i} = a_0 + a_1 Z_i + d_i \quad (2)$$

$$b_{1i} = c_0 + c_1 Z_i + f_i \quad (3)$$

There are two second-step regression equations, the first of which treats the first-step intercepts as a function of the Z variable and the second of which treats the first-step regression coefficients as a function of Z. The parameters in Equations 2 and 3 are as follows:

a_0: the response on Y for persons scoring zero on both X and Z

a_1: the effect of Z on the average response on Y

c_0: the effect of X on Y for groups scoring zero on Z

c_1: the effect of Z on the effect of X on Y.

For the intercepts (b_{0i}, a_0, and c_0) to be interpretable, both X and Z must be scaled so that either zero is meaningful or the mean of the variable is subtracted from each score. In the example used here, X is continuous and centered around its mean and Z is an effect-coded (−1, 1) categorical variable. With this coding scheme, zero can be thought of as an "average" across the two types of groups (democratic and autocratic). The top of Table 12 gives the interpretation of these four parameters for our example.

To repeat, there are two steps in the estimation procedure. In the first step, the slope and the intercept, b_{1i} and b_{0i}, are estimated for each group. That is, for each group Y is regressed on X as in Equation 1 and the slope b_{1i}, and intercept b_{0i}, are estimated. In the second step these slopes and intercepts serve as criterion measures in two regression equations in which Z is the predictor variable.

This two-step analysis procedure, where regression slopes from the first step are dependent variables in the second step, may seem unfamiliar. In fact, this procedure is similar in important respects to a conventional regression analysis with interactions between the lower-level predictor X and the upper-level predictor Z. To see this clearly, we need to substitute the terms for b_{0i} and b_{1i} in Equations 2 and 3 into Equation 1. This results in the following combined equation:

$$Y_{ij} = a_0 + a_1 Z_i + c_0 X_{ij} + c_1 Z_i X_{ij} + d_i + f_i X_{ij} + e_{ij} \quad (4)$$

If we ignore the terms d_i and $f_i X_{ij}$, we can see that this equation is identical to a conventional regression equation that specifies an interaction between X and Z; i.e., the effect of the lower-level predictor X depends on the upper-level predictor Z. A conventional regression model of this sort contains only one random effect, e_{ij}. However, the multilevel model specified in Equation 4 contains two additional random effects: d_i, a random intercept effect; and $f_i X_{ij}$, a random slope effect for X_{ij}. It is essential that these

TABLE 12

Definition of Effects and Variance Components for the Example

Effect Estimate

Constant: a_0
Typical level of satisfaction across all recruits and platoons

Leadership Style (Z): a_1
Degree to which recruits in the autocratic condition are more satisfied than recruits in the democratic condition

Authoritarianism (X): c_0
Degree to which a recruit's authoritarianism relates to satisfaction, controlling for leadership style

X by Z: c_1
Degree to which the effect of leadership style varies by authoritarianism

Variance

Platoon: s_d^2
Platoon differences in the recruits' satisfaction, controlling for leadership style and level of authoritarianism

X by Platoon: s_f^2
Platoon differences in the relationship between authoritarianism and satisfaction, controlling for leadership style

Error: s_e^2
Within-platoon variation in satisfaction, controlling for authoritarianism

additional random effects be considered to draw correct inferences from multilevel data.

Table 13 presents the estimates from these second-step regressions for the fictitious platoon data set. They are presented under the column designated OLS (ordinary least squares). We see that the intercept is about four and a half units. The estimates in Table 13 indicate that recruits in groups with autocratic leaders say that they are more satisfied than recruits with democratic leaders by about two-thirds of a point. (Because of effect coding, the coefficient must be doubled.) Also, recruits high in authoritarianism are more satisfied than those low in authoritarianism, but the effect is only marginally significant. There was no significant evidence of a differential effect of authoritarianism on satisfaction for the two leadership conditions.

Weighted Regressions

The first-step regression coefficients are likely to differ in their precision. Some are estimated more precisely than others because they are based on more observations and because X varies more. It seems reasonable to weight the second-step equation by the precision of the first-step estimates. That is, groups whose coefficients are better estimated should count more than groups whose coefficients are not well estimated.

Although weighting greatly complicates multilevel analysis, it provides two important dividends. First, when the second step regressions are weighted, the estimates are more precise; i.e., they are, in principle, closer to the population values than the unweighted estimates. Second, because weighting corrects for sampling error, variances of effects can be estimated. For instance, it can be determined whether the effect of X on Y varies across upper-level units. In terms of the example, we can estimate the degree to which the relationship between authoritarianism and satisfaction varies across platoons. So weighting at step two provides important benefits. Two different weighting strategies are considered: weighted least squares (WLS) and maximum likelihood (ML).

WLS Weighting To determine the weight or the accuracy of each group's regression coefficient w_i, we use the sum of squares for X or SS_i. This weight depends on two factors: the number of observations for group i and the variance of X for group i. Note that the larger the value of w_i, the more precisely b_{1i} is estimated and the more it is weighted in the second-step regression equation.

To better understand the difference between OLS and WLS, Table 14 presents information concerning the relationship between authoritarianism and satisfaction for five of the twenty-three groups. Platoon nine's data should be weighted more heavily because that platoon has more recruits than platoon four. Note also that the weight of platoon eleven is larger than platoon nine's, despite the fact that platoon nine has more recruits than platoon eleven. The pattern in the weights happens because platoon nine has less variation in authoritarianism than platoon eleven and so platoon nine has a smaller weight.

Table 13 presents the WLS estimates which are somewhat similar to the OLS estimates. Platoons with autocratic leaders have more satisfied recruits, and recruits higher in authoritarianism are more satisfied. The interaction is now statistically significant. Recruits higher in authoritarianism are especially satisfied when they have autocratic leaders.

For the weighted solution, there are two variances of effects. First, there is variance in the intercepts of σ_d^2. (From Equation 2, d is the residual term in the second-step regression of the intercepts.) It measures the extent to which there are differences between upper-level units in their average scores on Y when X is zero, after removing variation due to Z. So σ_d^2 measures the degree to which platoons differ in average levels of satisfaction controlling for both the levels of authoritarianism in the platoon and leadership style. The interpretation of σ_d^2 depends on the meaning of zero for the X variable. Recall that the inter-

TABLE 13

Estimates and Tests of Coefficients and Variance Components for the Example

	Coefficients					
	OLS		WLS		ML	
	b	*t*	*b*	*t*	*b*	*t*
Constant	4.645	37.63**	4.598	39.76**	4.604	41.79**
Leadership (*Z*)	0.352	2.85*	0.346	2.99**	0.318	2.88*
Author. (*X*)	0.162	1.94†	0.206	2.58*	0.162	2.03*
X by *Z*	0.141	1.69	0.158	1.97†	0.163	2.04†
			Variances			
			s^2	*F*	s^2	χ^2/df
Platoon (*G/Z* or *d*)			0.109	1.55†	0.085	1.54†
X by *G/Z* (*f*)			0.087	2.42**	0.087	2.44**
Error			1.470		1.481	

**$p < .01$
*$p < .05$
†$p < .10$

cept for group *i* is the predicted score when *X* is zero. Because authoritarianism is centered on its mean, the intercept is the level of satisfaction for a recruit who is average in authoritarianism.

Additionally, there is variance in the coefficients or σ_f^2 (see Equation 3). This variance measures the extent to which the relationship between *X* and *Y* varies across the upper-level units after removing variation due to *Z*. So for the example, this variance assesses the degree to which the relationship between satisfaction and authoritarianism varies across platoons, controlling for leadership style.

To summarize, there are two group effects that are represented by variances:[6]

σ_d^2: group or upper-level differences in the average response controlling for *X* and *Z*

σ_f^2: group or upper-level differences in the effect of *X* on *Y* controlling for *Z*.

The bottom of Table 12 presents the interpretation of these variances for the platoon example.

To test whether there is significant variance in the first-

TABLE 14

Five Selected Platoons from the Example

Platoon	Group Size	Variance in Authoritarianism	Effect of Authoritarianism	Weight*
4	5	1.30	0.115	5.20
9	9	2.44	0.341	19.56
11	8	5.56	0.608	38.88
15	13	2.42	–0.489	29.08
18	9	3.61	–0.212	28.89

*Weight equals the variance of authoritarianism times the group size less one.

step regression coefficients, one must first compute the following:

s_f^2: the error variance from the weighted second-step regression where b_{1i} was treated as the criterion variable

s_p^2: the pooled error variance or $\Sigma(df_i s_{ei}^2)/\Sigma df_i$

where s_{ei}^2 and df_i are the error variance in group i and its degree of freedom. By pooling the error variances, we are assuming that they are homogeneous across the upper-level units (groups). The test of whether the first-step regression coefficients vary significantly ($\sigma_f^2 = 0$) is $F(k-2, N-2k) = s_f^2/s_p^2$ where k equals the number of groups and N the total number of observations in the data set. The estimate of the variance of the coefficients or σ_f^2 is

$$\frac{s_f^2 - s_p^2}{q} \tag{5}$$

The value of q can be viewed as an average of the weights. When the weights do not vary, q equals that weight. But when the weights do change, q is a quite complicated average of weights (Kashy, 1991; Kenny, Bolger & Kashy, 1997).

In the bottom of Table 13, for the platoon example, the variances are presented. The standard deviation for the platoons is about 0.3 and is only marginally significant. However, the effect of authoritarianism on satisfaction does vary significantly from platoon to platoon. Assuming a normal distribution of slopes, the results indicate that about 68 percent of the platoons have authoritarianism slopes between 0.206 ± 0.295. So, although in most platoons recruits higher in authoritarianism are more satisfied, in some platoons recruits who are lower in authoritarianism are more satisfied.

The value of q for the example is 23.900 for the slopes and 7.371 for the intercepts. As has been stated, the value of q can be viewed as an average weight. This WLS estimation method can be implemented using the GLM procedure done within SAS (Kenny, Bolger & Kashy, 1997).

Maximum Likelihood (ML) Weighting In multilevel modeling, WLS estimation is rarely used, the more common method being maximum likelihood. To explain the difference between ML and WLS weights, consider the simplest multilevel model, one with no X or Z variables. As an example, fifteen members of three platoons rate their satisfaction with the group.

The model is the familiar one-way ANOVA model with random effects. There is only one fixed effect in this model, the constant or typical level of satisfaction. At issue here is how to estimate that effect. There are two different approaches: the weighted mean (sum of all the observations divided by N) and the unweighted mean (sum of all the group means divided by k). The OLS estimate of the intercept is the unweighted mean whereas the WLS estimate is the weighted mean.

ML uses a compromise between these two means based on the ratio of the variation within each group's data and the variation between groups. Consider the two sets of data in Table 15. In both sets there are three platoons whose means are 6, 7, and 8. The unweighted mean for both data sets is 7.000 and the weighted mean is 6.733. The difference between the two data sets is that for data set A, there is no within-group variance (recruits from the same platoon agree) whereas for B there is considerable variation. The unweighted mean is the appropriate measure of the intercept for data set A. Because there is no variation within platoons, having more observations really does not increase the precision in the estimation. But for data set B, there is considerably more variation and so having more observations is meaningful. So for data set B, the ML estimate (using the computer program HLM; see below) of the intercept is 6.746, which is closer to the weighted mean than the unweighted mean. Thus, the ML estimate does not use either the weighted or unweighted mean but rather uses an appropriate compromise based on the data.

The ML weight used to estimate σ_d^2 is $s_d^2 + w_i s_p^2$ where w_i is the WLS weight for unit i, s_d^2 is the variance of the intercepts, and s_p^2 is the pooled error variance. Note that if there is no within-group variation ($s_p^2 = 0$), which is the case for the data set A of Table 15, then the weights are homogeneous and the unweighted mean is used. If, however, s_d^2 is near zero (its estimate for data set B is only 0.054), then the weighted mean is used.

Although there is an impressive statistical logic to ML weighting, it presents an estimation difficulty. To estimate σ_d^2 one must know s_d^2 beforehand because s_d^2 is used in the weighting. This is the fundamental computational difficulty with ML. It results in the following consequences: iterative solutions and approximate standard errors. For more extended descriptions of maximum likelihood estimation of multilevel models, the reader should consult Bryk and Raudenbush (1992); Hedeker, Gibbons, and Flay (1994); and Goldstein (1995).

There is now a wide array of specialized computer programs that calculate these maximum likelihood estimates: HLM/2L and HLM/3L (Bryk, Raudenbush, & Congdon,

TABLE 15

Illustration of Weighting

Group	Data Set A	Data Set B
1	6 6 6 6 6 6 6	3 4 5 6 7 8 9
2	7 7 7 7 7	4 6 7 8 10
3	8 8 8	6 8 10

Unweighted Mean = 7.000; Weighted Mean = 6.733.

1994), MIXREG (Hedeker, 1993), MLn (Goldstein, Rasbash, & Yang, 1994), as well as 5V within BMDP and PROC MIXED within SAS. It should be noted that these programs actually accomplish the estimation in one step (and then iterate) and do not take two steps as do the OLS or WLS approaches. Also, most programs implement a variety of estimation approaches, the most common being restricted maximum likelihood.

We used the computer program HLM (Bryk et al., 1994) to estimate the parameters for the example platoon data set. Both the OLS and WLS results are fairly similar to the ML estimates. HLM uses a chi square test to evaluate the statistical significance of the variances. To compare this value to the WLS *F* tests, we divided the chi square by its degrees of freedom. As with WLS, the test that the effect of authoritarianism varied across groups is statistically significant.

ML also provides a measure of the covariance between d_i and f_i, the degree to which group differences in the slopes and intercepts are correlated. (Although WLS does not estimate this term, WLS does not assume that it is zero.) For the example, the correlation between *d* and *f* is .338 which indicates that platoons with more satisfied recruits also had a more positive effect of authoritarianism on satisfaction.

Summary

There are three strategies for estimating models with group or upper-level as unit: OLS, WLS, and ML. An OLS solution is the simplest to accomplish, but least efficient, and there are no estimates of the variances of the effects. A WLS solution provides estimates of the variances, but it can be much less efficient than the ML method. The ML method is the most efficient and is the most computationally complex. However, multilevel software is becoming increasingly accessible. It seems certain that ML will become the method of choice for the estimation of multilevel models.

It is helpful to contrast multilevel modeling with ML estimation to ANOVA with least-squares estimation. The major differences between the two are presented in Table 16. Within multilevel modeling, variables are denoted as fixed or random. In ANOVA, such a distinction can be made, but it is usually not featured. There is typically only one random factor: person. However, in many settings there are multiple random variables. As has been extensively discussed in this chapter, in group research there are two random variables: person and group.

One potential benefit of the use of multilevel models in social psychology is that they are likely to promote greater awareness of the distinction between fixed and random effects. In particular, stimuli in social psychological experiments (e.g., targets, words, sentence-stems) should properly be treated as random effects, but instead they are treated as fixed. Strictly speaking, then, the results of such experiments do not generalize beyond the specific stimuli used. Persons, on the other hand, are always treated as random effects, and statistical theory permits generalization to the population from which they were drawn. It is ironic that we social psychologists, who theorize that the situation is more important than the person, use analysis methods that make the person more important than the situation.

Random factors are featured much more in multilevel modeling than are fixed factors. The analyst examines the variance due to the presumed random effects in the model. If such variances are near zero, the model would be reestimated with the term dropped. Also, fixed factors may interact with random factors.

Classically in ANOVA, designs are balanced. Equal sample sizes in each cell of the design, though not a requirement, are highly desirable. Although this assumption can sometimes be relaxed, ANOVA works best on designs that are balanced. Multilevel modeling can handle bal-

TABLE 16

Differences Between ANOVA and Multilevel Modeling Paradigms

	Paradigm	
Factor	**ANOVA**	**Multilevel modeling**
Random factors	One	More than one
Terms estimated	Effects	Effects and variances
Design	Balanced	Unbalanced
Missing data	None	Allowable
Estimation technique	Least squares	Maximum likelihood
Formulas for estimates	Yes	No

anced as well as unbalanced designs. Although multilevel models can handle unbalanced data, the principle of balance is important and researchers should strive for it when they design research. There is a comparable parallel in latent variable modeling. Although those models allow the researcher to analyze unreliable measures (see below), it is still desirable to have as reliable measures as possible.

A related feature of multilevel modeling is that it allows for missing data. Consider conventional repeated-measures designs. Within ANOVA, if there are missing data from a person on a repeated measure, that person would have to be dropped from the analysis or the missing data would have to be "imputed." Multilevel models are often able to analyze all the data. Another feature is the estimation method for multilevel models. ANOVA models are estimated by least squares and significance tests involve comparison of mean squares. Generally, multilevel modeling employs maximum likelihood estimation.

Within ANOVA one can use the raw data to estimate an effect, and very often the estimates are means. For many classic social psychological experiments, the entire study is captured by a 2 x 2 table of means. With multilevel modeling, very complicated algorithms are used and the estimates are so complicated that no formula can be used to provide an estimate. Instead an iterative algorithm is used to approximate an estimate. The estimates are not means but coefficients from a complicated two-level analysis. Thus, it is much more difficult to go from the raw data to the results from the statistical analysis.

Multilevel modeling is quite different from the standard ANOVA framework. It promises to allow for more efficient and more flexible estimation of models than ANOVA. Moreover, it is reasonable to expect that multilevel modeling will become easier to implement and to interpret. We return to these differences in the conclusion of the chapter.

IDENTIFICATION IN STRUCTURAL EQUATION MODELING

Contemporary research in social psychology routinely uses structural equation models (Breckler, 1990). These models employ latent constructs that are measured by imperfect indicators. The set of links between indicators and latent constructs is called the *measurement model* (see Judd & McClelland, 1998, in this *Handbook*), and the set of links between constructs (i.e., causal paths) is called the *structural model.*

Structural equation modeling involves four steps. In the first step, called *specification,* the researcher determines which indicators reflect which latent variable and what the causal relations between latent variables are. Specification not only involves stating what causes what, but also what does *not* cause what. Certainly all models are incorrectly specified. The goal is to specify a model that is not too complex that it cannot be estimated (see below), but not too simple that it is trivial. Such a blend of "complex simplicity" can be difficult to achieve.

In the second step, called *identification,* the researcher determines whether there is enough information to estimate the model. For many investigators, this is the most mysterious step. The focus of this section is to provide guidance on this topic.

In the third step, called *estimation,* the parameters of the model are estimated. For some models without latent variables (called *path analysis models*), multiple regression is the estimation method. The estimation of models with latent variables requires specialized structural equation computer programs and, generally, maximum likelihood estimation is the estimation method. Because maximum likelihood estimation of structural equation models usually presumes a multivariate normal distribution, less restrictive estimation methods have also been developed.

The statistical theory on which structural equation modeling is based presumes that the models are estimated using the covariance matrix (a covariance between two variables equals their correlation times the product of their standard deviations). In practice, many models do not require that the covariance matrix be used for estimation (Cudeck, 1989), and a correlation matrix can be used instead. It is possible to reformulate the statistical theory and presume that the model is estimated from a correlation, not a covariance, matrix (Browne & Mels, 1994).

In the fourth step, the fit of the model is evaluated. If the fit is poor, the model can be respecified and so part of the evaluation of model fit is the determination of where the poor fit lies. Since the pioneering paper by Bentler and Bonett (1980), literally hundreds of measures of model fit have been developed (Bollen & Long, 1993). Although the choice of model fit depends on a host of factors, the Tucker-Lewis or nonnormed measure (Bentler & Bonett, 1980) is very often quite informative (Marsh, Balla, & McDonald, 1988).

After all four steps, the model is usually respecified based on the analysis of the data. In structural equation modeling, the researcher usually cycles through the steps of specification, identification, estimation, and model fit many times. Models that are respecified based on the data are exploratory and not confirmatory. Generally, the significance testing within structural equation modeling presumes that the model was specified without looking at the data. Capitalization on chance is a serious problem when models are substantially altered based on the analysis of the data (MacCallum, Roznowski, & Necowitz, 1992). Exclusive reliance on statistical and not theoretical criteria for respecification can lead to misleading models.

In the estimation step, the measured variables are correlated, and their correlations (or more generally covariances) are used for parameter estimation. However, sometimes the information available from the correlations is not sufficient to enable the researcher to estimate the parameters. For instance, for a model in which two variables indicate a single latent variable, there is a single correlation. With that one correlation, it is impossible to solve for the two factor loadings. The problem is a standard algebraic one of fewer equations than unknowns. In cases such as this, there is no unique mathematical solution for the model's parameters and the model is said to be *not identified.* An essential task in structural equation modeling is to establish whether the model is identified.

Traditionally, the determination of a model's identification status requires a formal mathematical analysis that is described in texts like Bollen's (1989). In practice, this analysis is too complex for most researchers, and so they rely on structural equation modeling software to determine whether a given model is identified (Hayduk, 1987). However, finding out that one's model is not identified at the data analysis stage means that the study has to be redesigned and rerun. Researchers need to know if their models are identified, in principle, *before* they collect the data. This section presents rules so that researchers can determine if their models are identified before they are estimated.

However, even models that are in principle identified may not be identified when they are actually estimated. Such models are said to be *empirically underidentified.* A simple example can illustrate this condition: if a causal variable does not vary in the sample, its effects cannot be empirically estimated. So a second purpose of this section is to assist researchers in the recognition of models that are not identified empirically.

The following set of rules can be used to check whether a given model is identified. What follows should be taken as a guide and not as gospel. The rules, by no means exhaustive, are nonetheless helpful in determining identification. What follows presumes some knowledge of structural equation modeling. If the reader lacks that knowledge, this section should be skipped.

If both the structural and the measurement models are identified, then the entire model is identified. For the entire model to be identified, the structural model must be identified. Occasionally, the measurement model alone is not but the entire model may be identified when certain paths in the structural model are set to zero (see Condition B3b).

Measurement Model Identification

In the measurement model, indicators are used to assess constructs or latent variables. For example, ratings by three friends of a target's friendliness might serve as indicators of target friendliness, and ratings of intelligence from the same three friends might be indicators of intelligence. Pairs of latent constructs may be correlated. That is, there may be a correlation between the constructs of friendliness and intelligence. Variance in the indicators that is not due to the latent constructs is called measurement error, and the errors of two indicators may be correlated. For example, the ratings by the same friend of friendliness and intelligence are likely correlated due to a halo effect.

For the measurement model to be identified, five conditions labeled A through E must hold. Conditions A and B must be satisfied by each construct, Condition C refers to each pair of constructs, and Condition D refers to each measure or indicator. Condition E refers to indicators that load on two or more constructs.

These rules primarily concern models in which each measure loads on only one construct. Fortunately, most estimated models in social psychology are of this type. If a variable loads on more than one construct, that variable is set aside and is discussed under Condition E.

Condition A: Scaling the Latent Variable Because a latent variable is unmeasured, its units of measurement must be fixed by the researcher. This condition concerns the manner in which the units of measurement are fixed. Each construct must have either:

1. one fixed nonzero loading (usually 1.0),
2. for causal or exogenous factors, fixed factor variance (usually 1.0), or for factors that are caused, fixed factor disturbance variance (usually 1.0), or
3. a fixed causal path (usually 1.0) leading into or out of the latent variable (see Kenny, 1979, pp. 180–182).

Some computer programs require that only strategy one be used, but the other two strategies are perfectly legitimate. For pure measurement model situations (no causation between latent variables) or confirmatory factor analysis, strategy two is often used, yielding the standard factor analysis model. Strategy three is hardly ever used.

Condition B: Sufficient Number of Indicators per Construct For each construct in the model, at least one of the following three conditions must hold.

1. The construct has at least three indicators whose errors are uncorrelated with each other.
2. The construct has at least two indicators whose errors are uncorrelated and either
   a. both the indicators of the construct correlate with a third indicator of another construct but the two indicators' errors are uncorrelated with the error of that third indicator (i.e., the two constructs must be correlated), or
   b. the two indicators' loadings are set equal to each other.
3. The construct has one indicator and either:
   a. the indicator's error variance is fixed to zero or some other a priori value (e.g., the quantity one minus the reliability times the indicator's variance), or

b. there is another variable that can serve as an instrumental variable (see Rule C under "Identification of the Structural Model" below) in the structural model and the error in the indicator is not correlated with that instrumental variable.

Condition C: Construct Correlations For every pair of constructs either

1. there are at least two indicators, one from each construct, that do not have correlated measurement error between them, or
2. the correlation between the pair of constructs is specified to be zero (or some other a priori value).

Condition D: Loading Estimation For every indicator, there must be at least one other indicator (not necessarily of the same construct) with which it does not share correlated measurement error. If the three above conditions hold and Condition D does not, then drop from the model all indicators that do not meet this condition and the model is still identified.

Condition E: Estimation of Double Loadings One important model in which *all* indicators have double loadings is the classic model for the multitrait-multimethod matrix (Campbell & Fiske, 1959). Each indicator loads on both a trait and method factor. Kenny and Kashy (1992) have shown that there are serious empirical identification difficulties with this model. All but the most adventurous researchers are well advised to avoid the estimation of such models. Alternative forms of multitrait-multimethod matrix models can be estimated (Kenny & Kashy, 1992; Millsap, 1995; Wothke, 1995).

However, a subset of indicators may load on two or more factors as long as Conditions A, B, and C are met for those constructs by including some indicators that load on only one construct. We refer to this as Condition E. Consider the indicator X_1 that loads on more than one construct. The errors of X_1 may be correlated with the errors of other indicators, but for each construct on which X_1 loads, there must be at least one singly-loading indicator that does not share any correlated error with X_1.

The rule in the previous paragraph is a sufficient condition for the identification of models with double-loading indicators. That is, some models that do not meet Condition E are identified in principle. However, in practice these models are often empirically under-identified.

Summary

For most measurement models, Condition E is not relevant, and it is usually very easy to verify that C and D are satisfied. Condition A can always be satisfied (but the researcher must make sure that it is), and so ordinarily the key condition to scrutinize carefully is B.

Constructs with a single error-free indicator (e.g., gender) are best handled by fixing their loading to one, forcing their error variance to zero, and leaving their variances free to be estimated. Of course, the assumption of zero error variance must be justified theoretically.

Empirical Identification of the Measurement Model Some models that are identified in theory are not identified for a particular study. Following Kenny (1979), these models are called *empirically underidentified models.* Consider a simple one-factor model with three indicators, X_1, X_2, and X_3. It can be shown that the standardized loading of X_1 on the factor equals the square root of $r_{12}r_{13}/r_{23}$. Mathematically for there to be a solution r_{23} must not equal zero. If its value is near zero, then there is no well-defined solution and the model is said to be empirically underidentified.[7]

Condition B, the number of indicators per construct, is critical to the empirical identification of each construct. Condition B1 requires three indicators. For these three indicators, each of the three correlations between those indicators should be statistically significant and the product of the three correlations must be positive.

Condition B2a has three indicators, two of which load on the latent variable and the third loads on another factor. As with B1, the three indicators must correlate significantly with each other and their product must be positive. If the two indicators of one construct correlate with the one indicator of the other construct, then those two constructs must be correlated.

For two indicators that are assumed to have equal loadings (Condition B2b), the correlation between the two must be significantly positive. If the correlation is large but negative, given theoretical justification, the loadings can be forced to be equal but of opposite signs.

If there is a single indicator and instrumental variable (Condition B3b) estimation is used, the indicator must share unique variance with the instrument (see Rule C below).

If the latent variable is scaled by fixing a loading to one (Condition A1), the indicator with a loading of one must correlate with other indicators of the latent variable. If all of the loadings are free and the disturbance of residual variance is fixed, empirical identification problems can occur if all or nearly all of the variance of the latent variable is explained by the other latent variables in the model.

If an indicator loads on two constructs (Condition E), the correlation between these two constructs cannot be very large. If that correlation is too large, the resulting multicollinearity makes it difficult to determine the indicator's loadings on the two factors.

Illustrations Figure 1 contains a series of examples. The latent factors or constructs are denoted by circled Fs and Gs, the measures are denoted by Xs, and the errors by Es. The reader should decide whether each model is identified and then read the text that follows.

Models I through IV are single-factor models. So Conditions C and E do not apply. Condition A is met in each model because one path is fixed to one. The key condition for these models is Condition B, but D must also be checked.

Model I meets Condition B1 and so is identified. Model II does not meet Condition B and so is not identified. Model III, though very similar to Model II, is identified because it meets Condition B2. Model IV meets Condition B1 but indicator X_4 fails to meet Condition D. So to identify this model, indicator X_4 must be dropped from the model.

Models V through VIII are two-factor models. All the conditions need to be checked, but we concentrate on Condition B. Models V and VI are not identified because they fail to meet Condition B. Both models would be identified if the errors were not correlated. Model VII is identical to Model VI, but there are three indicators per factor instead of two. This model is identified because it meets both Conditions B1 and B2a. Model VIII is identified because it meets Condition B2a. However, Model IX is not identified because B2a is not met since X_1 and X_2 do not correlate with X_3 and X_4 given that the correlation between the constructs is zero. Finally, Model X is identified. Condition B2a is met for both factors and so under Condition E, X_3 can load on both latent variables.

Identification of the Structural Model

The structural model consists of a set of causal equations. Variables that serve only as causes in the model are called *exogenous variables*. Unexplained variation in the effect variable is referred to as *disturbance*.

Rule A: Minimum Condition of Identifiability Let k be the number of constructs in the structural model and q be equal to $k(k - 1)/2$. The *minimum condition of identifiability* is that q must be greater than or equal to p where p equals the sum of:

a. the number of paths,
b. the number of correlations between exogenous variables,
c. the number of correlations between a disturbance and an exogenous variable, and
d. the number of correlations between disturbances.

In nearly all models, c is zero, and in many models d is zero. Theory places restrictions on a. Generally, b should be set at the maximum value; that is, all pairs of exogenous variables should be correlated.

If a structural model satisfies this minimum condition, the model *may* be identified. If it does not, the model is not identified; however, some but not all of the parameters of the model may be identified.

Rule B: Apparent Necessary Condition All models that satisfy the following condition appear to be identified: if between any pair of constructs, X and Y, no more than one of the following is true:

X directly causes Y

Y directly causes X

X and Y have correlated disturbances or if either X or Y is exogenous, it is correlated with the other's disturbance

X and Y are correlated exogenous variables.

Models that can be estimated by multiple regression form an important special case of this rule. For such models, the structural equations can be ordered such that if a variable appears as a cause in a given equation, it never later appears as an effect. Although we know of no proof for Rule B, we know of no exception. It seems likely that the rule generally holds.

Rule C: Instrumental Variable Estimation This rule considers exceptions to the previous rule: models that fail to satisfy Rule B but are nonetheless identified. The material in this section is very dense and may have to be read more than once. Because of these complications, this estimation method has rarely been used in social psychology, however, there have been some important uses of the method (e.g., Felson, 1981; Smith, 1982).

The estimation method is called *instrumental variable estimation*. An instrumental variable is assumed not to cause directly the effect variable. The absence of a causal path is what permits the estimation of an otherwise not identified model.

Consider X as a causal variable and Y as an effect variable. Instrumental variable estimation can be applied to the three following conditions:

1. spuriousness: an unmeasured variable causes both X and Y,
2. reverse causation: Y causes X, and
3. measurement error: measurement error in X which has only a single indicator.

Notice that conditions 1 and 2 violate Rule B. For all three conditions, the path from X to Y cannot be estimated by traditional means.

These models can be identified by instrumental vari-

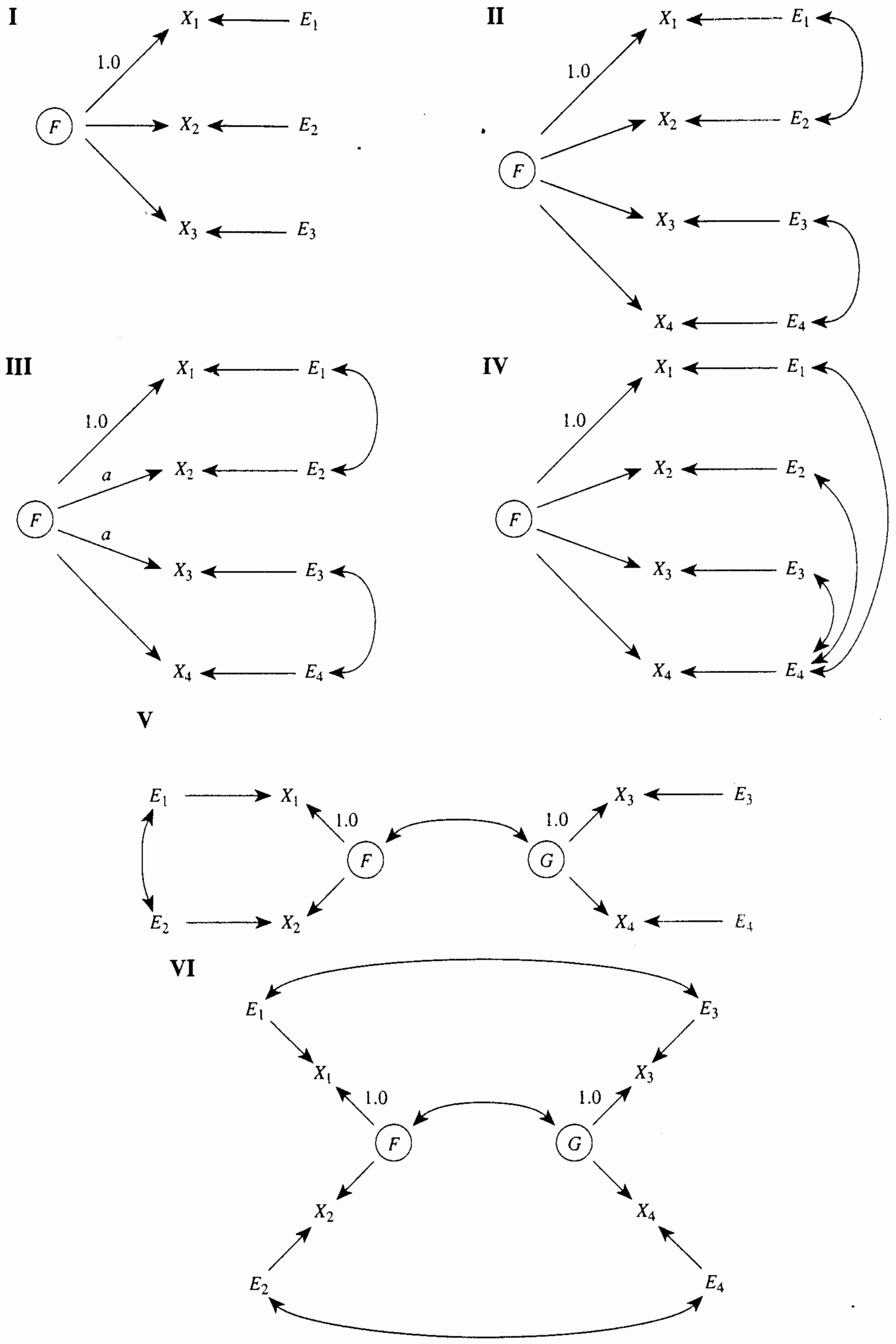
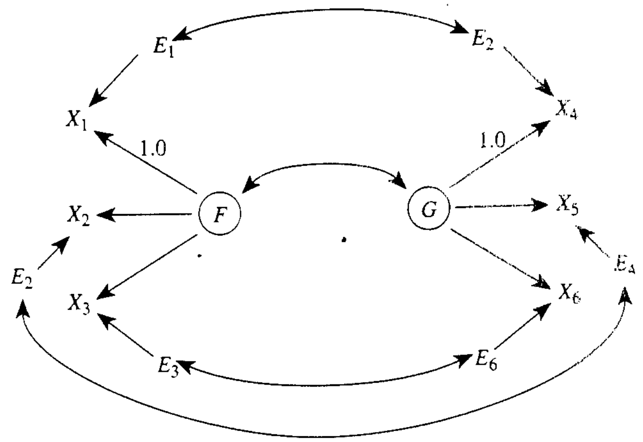
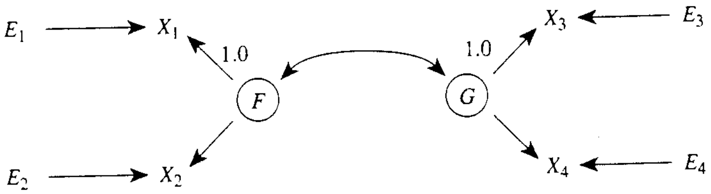


FIGURE 1 Measurement Model Identification Examples.
Latent constructs are denoted by circled *F*s and *G*s, measures by *X*s, and errors by *E*s.

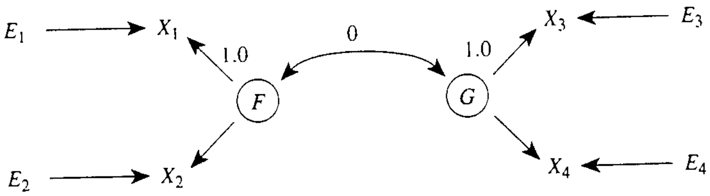
VII


E_1
E_2
X_1
X_4
1.0
1.0
X_2
F
G
X_5
E_2
E_4
X_3
X_6
E_3
E_6

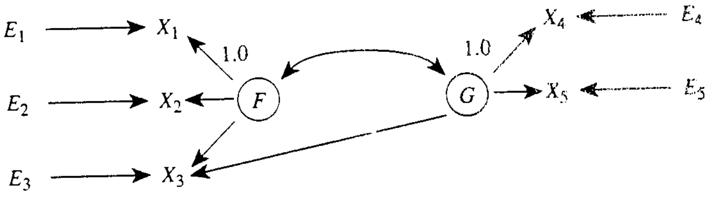
VIII


E_1
X_1
1.0
1.0
X_3
E_3
F
G
E_2
X_2
X_4
E_4

IX


E_1
X_1
0
1.0
1.0
X_3
E_3
F
G
E_2
X_2
X_4
E_4

X


X_4
E_4
E_1
X_1
1.0
1.0
E_2
X_2
F
G
X_5
E_5
E_3
X_3

ables. Denote X as a causal variable that meets one of the three above conditions, Y as the effect variable, U as Y's disturbance, I as an instrumental variable, and Z as a causal variable not needing an instrumental variable. The defining feature of an instrumental variable is that I is assumed not to cause Y directly: the path from I to Y is zero. This zero path is used to identify the model, but it must be given by theory, not by empirical analysis.

The following conditions are necessary for a model with instrumental variables to be identified.

1. The variable I must not directly cause Y or be correlated with U.
2. There must be as many or more I variables as there are X variables.[8]

Empirical Identification of the Structural Model If two variables that cause a third variable are strongly correlated, a condition called *multicollinearity*, the paths are not very precisely estimated. When there is perfect correlation between a pair of causal variables, the coefficients for those causal variables cannot be estimated at all.

For models with instrumental variables, the conditions for empirical identification are complicated. As before, let X be the variable that needs an instrumental variable, I be an instrumental variable, and Z be a causal variable that does not need an instrument. After the partialling out variance due to Z, the set of I variables must significantly correlate (i.e., have a large multiple correlation) with X. For there to be a correlation between I and X and for Rule C1 not to be violated, the following must hold for the appropriate use of instrumental variable estimation.

1. For spuriousness, X cannot cause I and I cannot be correlated with the omitted variable.
2. When X is measured with error, variable I cannot be correlated with the measurement error in X, however I itself may have measurement error. That part of X that contains error may cause I.
3. For a feedback relationship, X cannot cause I.

If there is more than one X variable (i.e., variables needing instruments) that cause Y, when each X is regressed on I and Z, the correlation between the predicted Xs should not be too large (see Kenny [1979] p. 91). Finally, in a feedback loop, the same variable cannot serve as the instrumental variable for both variables in the loop.[9]

Illustrations Figure 2 contains six models containing four variables, and the question is whether the structural model is identified or not. Model I is not identified because it fails to satisfy Rule A, the minimum condition of identifiability. There are four variables and so there are six correlations. Because there are eight parameters to be estimated in Model I, the minimum condition of identifiability has not been met. All the remaining models meet that condition.

The next three models are identified because they satisfy Rule B. Models II and III can be estimated by multiple regression, and a variant of Model III is presented in the section on mediation. Model IV contains a feedback cycle (X_1 causes X_2, X_2 causes X_3, X_3 causes X_4, and X_4 causes X_1), and it is identified.

Because Models V and VI do not meet Rule B, we need instrumental variables and Rule C to identify these models. Model V has a feedback loop between X_3 and X_4. The variable X_1 can serve as an instrumental variable in estimating the effect of X_4 on X_3, and X_2 can serve as instrumental variable in estimating the effect of X_3 on X_4. For Model VI, the path from X_2 to X_4 needs an instrumental variable, because X_2 is correlated with the disturbance in X_4. It would seem that X_3 could serve as an instrumental variable, but it cannot because X_2 causes X_3 making the instrument correlated with the disturbance in X_4. Given this correlation, X_3 cannot serve as an instrumental variable and so Model VI is not identified.

Conclusion

This section on identification is very dense, but it does address an important and neglected issue in causal modeling. To design intelligent measurement and the structural models, the researcher needs to know whether the models are identified. Waiting to find out that model is not identified during the estimation stage is too late. Issues of identification are particularly relevant for the testing of mediational models, the topic of the next section.

MEDIATIONAL ANALYSIS

Structural equation modeling greatly facilitates the estimation and testing of causal sequences, particularly those involving theoretical constructs rather than measured variables. One particular type of causal model, a model proposing a mediational process, occurs frequently in social psychology. Very often a phenomenon is discovered (e.g., social facilitation or group polarization), and researchers are eager to discover the process by which the phenomenon operates. As discussed by Taylor (1998, in this *Handbook*), much of what social psychologists do is attempt to understand how internal processes mediate the effect of the situation on behavior. When a mediational model involves latent constructs, structural equation modeling provides the basic data analysis strategy. If the mediational model involves only measured variables, the basic analytical approach is multiple regression. Regardless of which data-analytic method is used, the steps necessary for testing mediation are the same. In this section, we describe the analyses required for testing mediational hypotheses

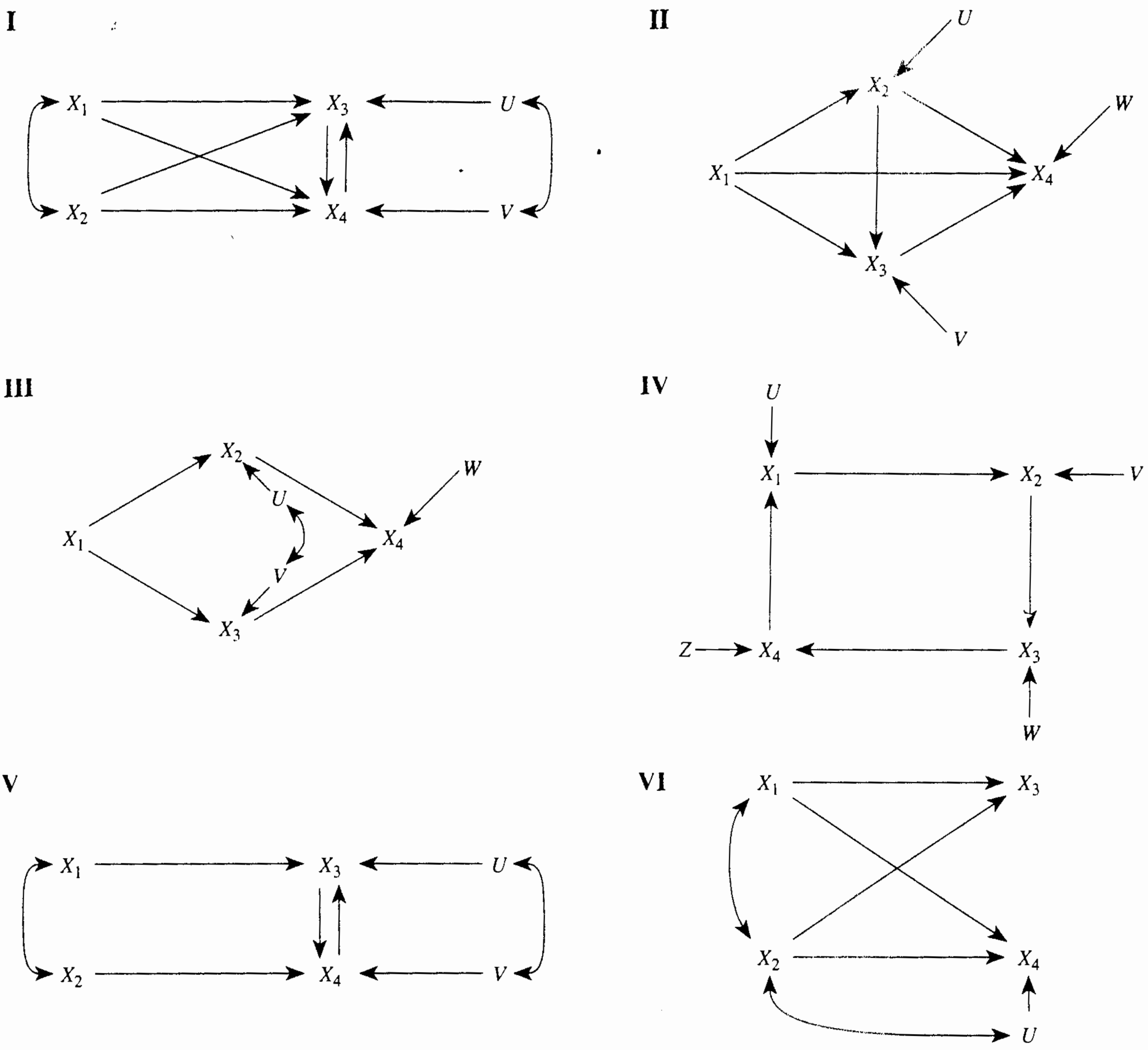


FIGURE 2 Structural Model Identification Examples, Each with Four Variables.

and we address several questions that such analyses have engendered.

Consider a variable X that is assumed to affect another variable Y. The variable X is called the *initial variable* and the variable that it causes or Y is called the *outcome*. In diagrammatic form, the unmediated model is presented in Figure 3(a). The effect of X on Y may be mediated by a process variable M, and the variable X may still affect Y. The mediated model is presented in diagrammatic form in Figure 3(b). The mediator has been called an *intervening* or *process* variable. *Complete mediation* is the case in which variable X no longer affects Y after M has been controlled and so path c' in Figure 3(b) is zero. *Partial mediation* is the case in which the path from X to Y is reduced in *absolute* size but is still different from zero when the mediator is controlled.

Baron and Kenny (1986) and Judd and Kenny (1981) have discussed four steps in establishing mediation.

Step 1. Show that the initial variable is correlated with the outcome. Use Y as the criterion variable in a regression equation and X as a predictor—estimate and test path c in Figure 3(a). This step establishes that there is an effect that may be mediated.

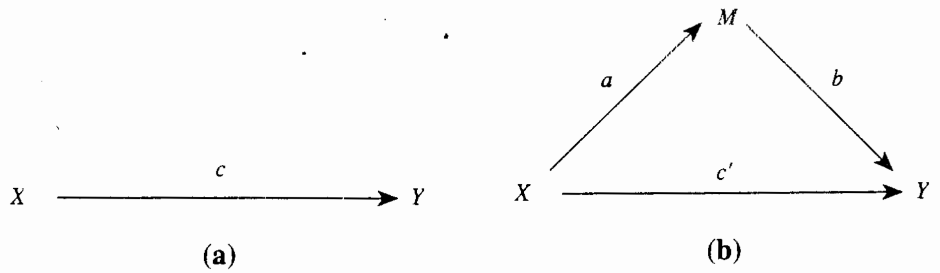


FIGURE 3 Basic Mediational Structure.
X is the initial variable, *Y* the outcome variable, and *M* the mediator.

Step 2. Show that the initial variable is correlated with the mediator. Use *M* as the criterion variable in the regression equation and *X* as a predictor—estimate and test path *a* in Figure 3(b). This step essentially involves treating the mediator as if it were an outcome variable.

Step 3. Show that the mediator affects the outcome variable. Use *Y* as the criterion variable in a regression equation and *X* and *M* as predictors—estimate and test path *b* in Figure 3(b). It is not sufficient just to correlate the mediator with the outcome; the mediator and the outcome may be correlated because they are both caused by the initial variable *X*. Thus, the initial variable must be controlled in establishing the effect of the mediator on the outcome.

Step 4. To establish that *M completely* mediates the *X-Y* relationship, the effect of *X* on *Y* controlling for *M* should be zero—estimate and test path *c*′ in the Figure 3(b). The effects in both Steps 3 and 4 are estimated in the same regression equation.

If all four of these steps are met, then the data are consistent with the hypothesis that *M* completely mediates the *X-Y* relationship, and if the first three steps are met but Step 4 is not, then partial mediation is indicated. Meeting these steps does not, however, conclusively establish that mediation has occurred because there are other (perhaps less plausible) models that are consistent with the data (MacCallum, Wegener, Uchino, & Fabrigar, 1993). Some of these models are considered later in this section.

The amount of mediation is defined as the reduction of the effect of the initial variation on the outcome or $c - c'$. This difference in coefficients can be shown to equal exactly the product of the effect of *X* on *M* times the effect of *M* on *Y* or *ab* and so $ab = c - c'$. Note that the amount of reduction in the effect of *X* on *Y* is not equivalent to either the change in variance explained or the change in an inferential statistic such as *F* or a *p* value. It is possible for the *F* from the initial variable to the outcome to decrease dramatically even when the mediator has no effect on the outcome.

If Step 2 (the test of *a*) and Step 3 (the test of *b*) are met, it follows that there necessarily is a reduction in the effect of *X* on *Y*. An indirect and approximate test that $ab = 0$ is to test that both *a* and *b* are zero (Steps 2 and 3). Baron and Kenny (1986) provide a direct test of *ab* which is a modification of a test originally proposed by Sobel (1982). It requires the standard error of *a* or s_a (which equals a/t_a where t_a is the *t* test of coefficient *a*) and the standard error of *b* or s_b. The standard error of *ab* can be shown to equal approximately the square root of $s_a{}^2 s_b{}^2 + b^2 s_a{}^2 + a^2 s_b{}^2$ and so under the null hypothesis that *ab* equals zero, the following

$$\frac{ab}{\sqrt{s_a{}^2 s_b{}^2 + b^2 s_a{}^2 + a^2 s_b{}^2}}$$

is approximately distributed as *Z*. Measures and tests of indirect effects are also available within many structural equation modeling programs.

One might ask whether all of the steps have to be met for there to be mediation. Certainly, Step 4 does not have to be met unless the expectation is for complete mediation. Moreover, Step 1 is not required, but a path from the initial variable to the outcome is implied if Steps 2 and 3 are met.[10] So the essential steps in establishing mediation are Steps 2 and 3.

Example

Morse, Calsyn, Allen, and Kenny (1994) examined the effect of an intervention that was designed to reduce the number of days homeless. The participants in this research were 109 homeless adults in a large Midwestern city, and the intervention was an intensive case management program. A total of 46 persons was randomly assigned to the intervention and the remaining 63 were assigned to a comparison group. The intervention serves as the initial variable and is dummy coded such that 1 is treated and 0 is

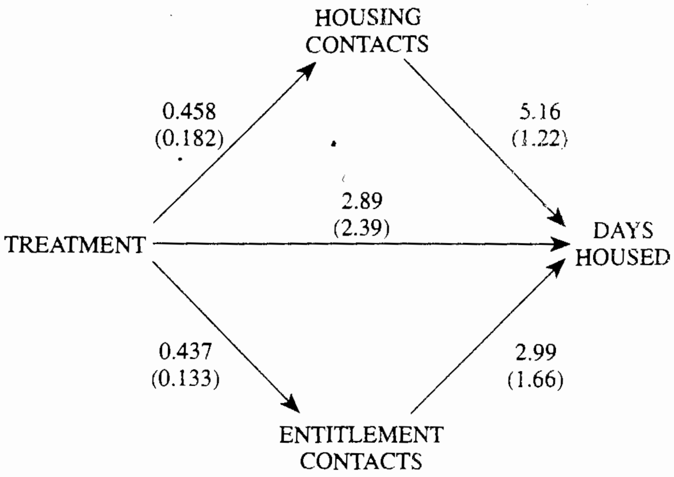


FIGURE 4 Mediation of Intervention Effects on Homelessness by Housing and Entitlement Contacts.

control. The outcome measure is the number of days housed per month during a period of 12 to 18 months after the intervention was initiated. For this illustration, two mediators are tested. The first mediator is the number of contacts per month about housing, and the second is the number of contacts per month about entitlements (money). Both mediators were measured for nine months after the intervention was initiated.

The total effect of the intervention on homelessness is 6.56 days, meaning that persons who received intensive case management were housed about one week more per month than those who did not. This effect is statistically significant ($p = .009$). So the Step 1 criterion is met.

Figure 4 presents the estimated coefficients and their standard errors in parentheses for this example. We see that the intervention resulted in both more housing and entitlement contacts. The effect for housing contacts is 0.458 ($p = .013$) and for entitlements it is 0.437 ($p < .001$). Both effects indicate thai individuals in the treatment group received an average of about five more contacts per year than did control group members. Because both effects are statistically significant, the Step 2 criterion is met for both mediators.

Also presented in Figure 4 are the effects from the mediators to the outcome. This effect is 5.16 for housing ($p < .001$) and 2.99 for entitlements ($p = .07$, not significant). The effect for the housing contact mediator indicates that for every monthly housing contact, the person was housed about five more days. The Step 3 criterion is met for only the housing contact mediator.

Finally, Figure 4 presents the effect of the intervention on the outcome, controlling for the mediators. That effect is now 2.89 ($p = .23$). Thus, the Step 4 criterion is met. It can then be concluded that housing contacts mediates the effect of the intervention on the outcome.

Because the unmediated effect was 6.56, 56 percent of the total effect is explained. The total reduction in the effect is 2.36 due to housing (0.458 times 5.16) and 1.31 due to entitlements (0.437 times 2.99). The sum of these two effects exactly equals the reduction in the effect of the intervention when the mediators are introduced. Using the Baron and Kenny modification of the Sobel test, the reduction due to housing contacts is statistically significant ($Z = 2.12, p = .034$) and the reduction due to entitlements is not ($Z = 1.52, p = .13$).

Problems in Testing Mediation

There are several issues that complicate the analysis of mediation. They can be divided into design issues, specification issues, and multilevel data. These issues are considered in turn.

Design Issues

Distal and Proximal Mediation To demonstrate mediation both paths a and b [see Figure 3(b)] need to be relatively large. Usually, the maximum size of the product ab is c, and so as a increases, b must decrease and vice versa.

The mediator can be too close in time or in the process to the initial variable and so a would be relatively large and b relatively small. An example of a *proximal* mediator is a manipulation check. The use of a proximal mediator may create multicollinearity which is discussed in the next part.

Alternatively, the mediator can be chosen too close to the outcome and with a *distal* mediator b is large and a is small. Ideally, standardized a and b should be comparable in size.

Multicollinearity If M is a successful mediator, it is necessarily correlated with X due to path a. This correlation, called collinearity, affects the precision of the estimates of the last set of regression equations. If X explains all the variance in M, then there is no unique variance in M to explain Y. The power of the tests of the coefficients b and c' is compromised. The effective sample size for these tests is approximately $N(1 - r_{XM}^2)$ where N is the total sample size and r_{XM} is the correlation between the initial variable and the mediator. So if M is a strong mediator (path a is large), to achieve equivalent power the sample size would have to be larger than what it would be if M were a weak mediator.

Specification Errors

Reverse Causal Effects The mediator may be caused by the outcome variable (Y would cause M in Figure 3). When the initial variable is a manipulated variable, it cannot be caused by either the mediator or the outcome. But because both mediator and the outcome variables are generally not manipulated variables, they may cause each other. Often if the mediator and the outcome variable were interchanged, the outcome would seem to "cause" the mediator.

Sometimes reverse causal effects can be ruled out theoretically. That is, a causal effect in one direction does not make sense. Design considerations may also weaken the plausibility of reverse causation. Ideally, the mediator should be measured before the outcome variable, as in the Morse et al. (1994) example.

If it can be assumed that c' is zero, then reverse causal effects can be estimated. (A review of the use of instrumental variables in the structural equation identification section may be helpful at this point.) That is, if it can be assumed that there is complete mediation (X does not directly cause Y), the mediator may cause the outcome and the outcome may cause the mediator.

Smith (1982) has developed another method for the estimation of reverse causal effects. Both the mediator and the outcome variables are treated as outcome variables, and they each may mediate the effect of the other. To be able to employ the Smith approach, for both the mediator and the outcome, there must be a different variable that is known to cause each of them but not the other. So a variable must be found that is known to cause the mediator but not the outcome and another variable that is known to cause the outcome but not the mediator. These variables are called *instrumental variables* (see the identification section).

Measurement Error in the Mediator If the mediator is measured with less than perfect reliability, then the effects are likely biased. The effect of the mediator on the outcome (path b) is likely underestimated and the effect of the initial variable on the outcome (path c') is likely overestimated if ab is positive (which is typical). The overestimation of c' is exacerbated to the extent to which the mediator is caused by the initial variable.

To remove the biasing effect of measurement error, multiple indicators of the mediator can be used to tap a latent variable. Alternatively, instrumental variable estimation can be used, but as before, it must be assumed that c' is zero. If neither of these approaches is used, the researcher needs to demonstrate that the reliability of the mediator is very high so that the bias is fairly minimal.

Omitted Variables This is the most difficult specification error to solve. The variance that the mediator shares with the outcome may be due to another variable that causes both the mediator and the outcome. Although there has been some work on the omitted variable problem (Mauro, 1990), the only complete solution is to specify and measure such variables and control for their effects.

Sometimes the source of covariance between the mediator and the outcome is a common method effect. Ideally, efforts should be made to ensure that the two variables do not share method effects (e.g., both are self-reports from the same person).

Multilevel Data and Mediational Analyses With multilevel data, there are two levels; for example, person may be the upper-level unit and time or day the lower-level unit. There are two types of mediation within multilevel models: the initial variable can be either an upper-level variable or a lower-level variable, but for both cases, the mediator and the outcome are lower-level variables. Before reading this part, the reader should have read the earlier section on multilevel modeling.

Upper-level Mediation Consider the effect of stress on mood. On each day for two weeks, stress and mood are measured for each person. Imagine that half the sample is classified as high on neuroticism and the other half is not. So the mediational question is the extent to which stress mediates the effect of neuroticism on mood.

This data structure naturally lends itself to an investigation of both mediation and moderation (Baron & Kenny, 1986). First, there is the simple mediational hypothesis: those high on neuroticism experience more stress and that stress leads to negative moods. Second, there is the moderation hypothesis that those high on neuroticism may react to stress more than those who are not. So stress might moderate the effect of neuroticism on mood. Following Bolger and Schilling (1991), the total effect of neuroticism on mood can be partitioned into a mediation and a moderation piece.[11]

With multilevel modeling of over-time data, the general hypothesis that person interacts with a within-person variable can be tested. For the example, within levels of neuroticism, stress may have more or less of an effect on mood for some individuals than for others, i.e., some individuals react more to stress than do others. Multilevel modeling provides a method for determining whether processes vary by person.

Lower-level Mediation We modify the example by making stress the initial variable and adding the lower-level variable, coping, that is assumed to mediate the stress-mood relationship, a model studied by Bolger and Zuckerman (1995). So coping, a lower-level variable, is triggered by stress and coping then elevates mood. It can be tested whether the mediational effects of coping vary across persons. First, the effect of stress could lead to coping for some persons and not for others. Second, coping may improve the mood of some but not others. In this way we can discern how the mediation of coping effects are moderated by individual differences.

Both mediational effects (the effect of stress on coping and the effect of coping on mood) can be treated as dependent variables, and measures of individual differences can be used to predict them. So, for example, it can be tested whether people who receive training in coping with stress do in fact use coping strategies when they experience stress and whether this coping is effective in raising mood. The training variable would serve as a moderator of the mediational process.

SUMMARY

Data analysis in social psychology all too often is a mindless exercise. Data are gathered using a standard design (a factorial experiment) and the same statistical analysis is performed (ANOVA followed by post hoc tests of means). An anthropologist might describe social psychological data analysis as a ritualistic exercise with Greek incantations mixed with practices developed in the early twentieth century to test the relative advantages of crop fertilizers.

Data analysis should be a more thoughtful process. Careful consideration should be given to the process that generated the data. Even with a factorial experiment, attention must be given to model assumptions. We have emphasized the assumption of independence in the first two sections of the chapter, but the other assumptions merit scrutiny (Judd et al., 1995; Wilcox, 1987).

We must learn to model a process not just to analyze data. Mediational analyses are likely to be helpful toward meeting this aim. Generally mediational analyses require structural equation modeling. This chapter has provided a detailed analysis of mediation, as well as provided advice concerning the difficult issue of identification. Moreover, we have discussed mediational analyses of multilevel data.

In the last twenty years we have witnessed a paradigm shift in the analysis of correlational data. Confirmatory factor analysis and structural equation modeling have replaced exploratory factor analysis and multiple regression as the standard methods. We are currently in the early stages of a paradigm shift in the analysis of experimental data. Multilevel modeling is replacing ANOVA. Certainly, ANOVA will remain a basic tool in social psychological research, but it can no longer be considered the only technique. Many models can be more efficiently estimated by multilevel modeling than by ANOVA. More importantly, many scientifically interesting hypotheses can be tested within multilevel modeling that cannot be easily addressed within an ANOVA framework.

Multilevel modeling is the wave of the future, and social psychology must begin to use it in research or other disciplines may lay claim to more of what is traditionally viewed as social psychology. Many traditional social psychological topics, such as group behavior and close relationships, are now being studied more by our colleagues in communications, family studies, and organizational behavior. If we continue to conceptualize social psychological research in terms of 2 × 2 designs and ANOVA, we will further narrow the scope of our field.

NOTES

1. Some meta-analysts are engaging in problematic practices. Occasionally, p values are reported as one-tailed tests when they should be two-tailed. Also very often large-sample theory is used (e.g., a Z test) when there is an available small-sample test (e.g., a t test). Finally, univariate tests (e.g., r or t tests of means) are used when multivariate tests (e.g., multiple regression) should be.
2. Some of the material in this section parallels the discussion of Crits-Christoph and Mintz (1991). Computations based on the formulas in Table 3 of this chapter were able to reproduce many of Crits-Christoph and Mintz's simulation results.
3. If a value of .10 seems too large, it should be realized that the effective alpha for the study would be only .06 because 80 percent of the time the nonindependence would be detected resulting in an alpha of .05 and 20 percent of the time the effective alpha would be .10.
4. Because only positive values of the intraclass correlation make the significance test of treatment effects too liberal (see Table 4), only positive values of the intraclass correlation are tested and so the tests in Table 6 are one-tailed.
5. The intraclass correlations, ρ_G and ρ_{GxA}, when computed using the formulas in Tables 1 and 4, are actually partial correlations. That is, the variance due to G is removed when ρ_{GxA} is computed, and the variance due to GxA is removed when ρ_G is computed. Sometimes the regular, nonpartial correlations are needed, and they are presented in Table 3 as ρ_1 and ρ_2.
6. Though not obvious, σ_d^2 takes on the role of ρ_G (discussed in the unit of analysis section) and σ_f^2 takes on the role of ρ_{GxA}.

7. Note that when r_{23} is zero, the estimate of the loading equals infinity. In physics, a black hole occurs when a particular equation has zero in its denominator. So empirical underidentification is causal modeling's equivalent of a black hole.
8. If an instrumental variable is needed because X has measurement error, then X need have an instrument in only one equation in which it is a causal variable. However, for both spuriousness and feedback, X needs to have an instrumental variable each time such conditions arise.
9. Alternatively one variable in the feedback loop need not have an instrument if the disturbances of the two variables in the loop are uncorrelated.
10. If c' is opposite in sign to ab, then it could be the case that Step 1 is not met, but there is still mediation. In this case the mediator acts like a suppressor variable.
11. Because of the differential weighting of estimators across groups that occurs in multilevel modeling, the total effect does not usually exactly equal the sum of the direct and the mediated or indirect effects.

REFERENCES

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Erlbaum.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6,* 267–285.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Bolger, N., & Schilling, E. A. (1991). Personality and the problems of everyday life: The role of neuroticism in exposure and reactivity to daily stressors. [Special Issue: Personality and daily experience.] *Journal of Personality, 59,* 355–386.

Bolger, N., & Zuckerman, A. (1995). A framework for studying personality in the stress process. *Journal of Personality and Social Psychology, 69,* 890–902.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bollen, K. A., & Long, J. S. (Eds.) (1993). *Testing structural equation models.* Newbury Park, CA: Sage.

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin, 107,* 260–273.

Browne, M. W., & Mels, G. (1994). *RAMONA: User's guide.* Ohio State University, Psychology Department.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1994). *Hierarchical linear modeling with the HLM/2L and HLM/3L programs.* Chicago, IL: Scientific Software International.

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows.* Newbury Park, CA: Sage.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cooper, H., & Hedges, L. V. (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Crits-Christoph, P., & Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of pychotherapies. *Journal of Consulting and Clinical Psychology, 59,* 20–26.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105,* 317–327.

Felson, R. B. (1981). Self- and reflected appraisal among football players: A test of the Meadian hypothesis. *Social Psychology Quarterly, 44,* 116–126.

Goldstein, H. (1995). *Multilevel statistical models.* New York: Halstead Press.

Goldstein, H., Rasbash, J., & Yang, M. (1994). *MLn: User's guide for Version 2.3.* London: Institute of Education, University of London.

Griffin, D., & Gonzalez, R. (1995). Correlational analysis of dyad-level data in the exchangeable case. *Psychological Bulletin, 118,* 430–439.

Harris, R. J. (1994). *ANOVA: An analysis of variance primer.* Itasca, IL: Peacock.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances.* Baltimore: Johns Hopkins Press.

Hedeker, D. (1993). *MIXREG. A FORTRAN program for mixed-effects linear regression models.* Chicago, IL: University of Illinois.

Hedeker, D., Gibbons, R. D., & Flay, B. R. (1994). Random-effects regression models for clustered data with an example from smoking prevention research. *Journal of Consulting and Clinical Psychology, 62,* 757–765.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5,* 602–619.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model-comparison approach.* San Diego: Harcourt Brace Jovanovich.

Judd, C. M., & McClelland, G. H. (1998). Measurement. In D. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 180–232). New York: McGraw-Hill.

Judd, C. M., McClelland, G. H., & Culhane, S. E. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. *Annual Review of Psychology, 46,* 433–465.

Jussim, L. (1991). Social perception and social reality: A reflection-construction model. *Psychological Review, 98,* 54–73.

Kashy, D. A. (1991). *Levels of analysis of social-interaction diaries: Separating the effects of person, partner, day, and interaction* (Ph.D. dissertation, University of Connecticut).

Kashy, D. A., & Kenny, D. A. (1997). The analysis of data from dyads and groups. In H. Reis & C. M. Judd (Eds.), *Handbook of research methods in social psychology.* New York: Cambridge University Press.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15, pp. 192–241). Lincoln: University of Nebraska.

Kelley, H. H., & Thibaut, J. W. (1978). *Interpersonal relations: A theory of independence.* New York: Wiley.

Kenny, D. A, (1979). *Correlation and causality.* New York: Wiley-Interscience.

Kenny, D. A. (1985). Quantitative methods for social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (3rd ed., Vol. 1, pp. 487–508). New York: Random House.

Kenny, D. A. (1995). The effect of nonindependence on significance testing in dyadic research. *Personal Relationships, 2,* 67–75.

Kenny, D. A. (1996a). The design and analysis of social interaction research. *Annual Review of Psychology, 47,* 59–86.

Kenny, D. A. (1996b). Models of independence in dyadic research. *Journal of Social and Personal Relationships, 13,* 279–294.

Kenny, D. A., Bolger, N., & Kashy, D. A. (1997). *Estimation of multilevel models using weighted least squares.* Unpublished paper, University of Connecticut.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin, 99,* 422–431.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112,* 165–172.

Kenny, D. A., & La Voie, L. (1985). Separating individual and group effects. *Journal of Personality and Social Psychology, 48,* 339–348.

Koele, P. (1982). Calculating power in analysis of variance. *Psychological Bulletin, 92,* 513–516.

Kraemer, H. C., & Jacklin, C. N. (1979). Statistical analysis of dyadic social behavior. *Psychological Bulletin, 86,* 217–224.

Latané, B. (1981). The psychology of social impact. *American Psychologist, 36,* 343–356.

Loehlin, J. C. (1992). *Latent variable models: An introduction to factor, path, and structural analysis.* Hillsdale, NJ: Erlbaum.

MacCallum, R. C., Roznowski, M., & Necowtiz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization of chance. *Psychological Bulletin, 111,* 490–504.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114,* 185–199.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

Mauro, R. (1990). Understanding L.O.V.E. (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin, 108,* 314–329.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective.* Belmont, CA: Wadsworth.

Millsap, R. E (1995). The statistical analysis of method effects in multitrait-multimethod data: A review. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 93–109). Hillsdale, NJ: Erlbaum.

Morse, G. A., Calsyn, R. J., Allen, G., & Kenny, D. A. (1994). Helping homeless mentally ill people: What variables mediate and moderate program effects? *American Journal of Community Psychology, 22,* 661–683.

Myers, J. L. (1979). *Fundamentals of experimental design* (3rd ed.). Boston: Allyn & Bacon.

Severo, N. C., & Zelen, M. (1960). Normal approximation to the chi-square and non-central *F* probability functions. *Biometrika, 47,* 411–416.

Smith, E. R. (1982). Beliefs, attributions, and evaluations: Nonhierarchical models of mediation in social cognition. *Journal of Personality and Social Psychology, 43,* 248–259.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural models. In S. Leinhardt (Ed.), *Sociological methodology 1982* (pp. 290–312). San Francisco: Jossey-Bass.

Taylor, S. E. (1998). The Social Being in Social Psychology. In D. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 58–95). New York: McGraw-Hill.

Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 181–227). San Diego: Academic Press.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications.* Cambridge, England: Cambridge University Press.

Wilcox, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology, 38,* 29–60.

Wothke, W. (1995). Covariance components analysis of the multitrait-multimethod matrix. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 125–144). Hillsdale, NJ: Erlbaum.